FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Riitta Rankinen

FORECASTING WIND ENERGY FOR A DATA CENTER

Master's Thesis
Degree Programme in Computer Science and Engineering
June 2021

Rankinen R. (2021) Forecasting wind energy for a data center. University of Oulu, Degree Programme in Computer Science and Engineering. Master's Thesis, 74 p.

ABSTRACT

Data centers are increasingly using renewables such as wind and solar energy. RISE's ICE data center has already solar panels and is now studying impact of adding a wind turbine into their microgrid. In this thesis, a machine learning model was developed to forecast wind power production for the data center.

Data center in Luleå has several applications to utilize wind power forecasting. Renewable energy sources are intermittent, so accurate forecasting of output power reduces a need for additional balancing of energy and reserve power in an electricity grid. Renewable energy can be reserved from market for next hour or next day to maximize its use. Forecasting from 30 min to 6 hours ahead allows job scheduling to optimize usage of renewables and to reduce power consumption. Data center may target to minimize electricity cost or maximize usage of renewables for lower greenhouse gas emissions. Smart microgrid based on artificial intelligence is the way to implement the applications.

Two open data sets from India and Sweden have been used in the research. The data available supports choosing of a statistical model. Random forest regression was the model used in the research. Data from India enabled to develop a model for one wind turbine. Developed model forecasted output power well. Swedish data set is from EEM20 competition, it included total wind power production in Sweden and had to be applied to approximate production of one wind turbine in Luleå. To achieve the goal output power of Luleå price region was averaged, and location for the simulation was chosen to be near Luleå. As expected, the accuracy of forecasting with Swedish data was reasonable, but approximations done reduced it.

The developed model was applied to RISE's ICE data center. Validation has been done, but final testing will take place in RISE's simulation environment. In general, data from northern Sweden is not openly available for wind power forecasting. In addition, any scientific articles covering the geographical area were not found while working on literature review. The study with Swedish competition data gave understanding, which variables are significant in northern Sweden and about their relative importances. Wind gust is such a variable. Using two data sets from different geographical locations proved that climate has a major impact on performance of the trained model. Thus, it is reasonable to use the trained model in locations with similar weather conditions only.

Keywords: wind energy, wind power, forecasting models, random forest regression, data center, renewable energy, sustainability

Rankinen R. (2021) Tuulienergian ennustaminen datakeskusta varten. Oulun yliopisto, tietotekniikan tutkinto-ohjelma. Diplomityö, 74 s.

TIIVISTELMÄ

Datakeskukset käyttävät uusiutuvia energialähteitä yhä enemmän. Tällaisia lähteitä ovat mm. tuuli- ja aurinkoenergia. RISE:n ICE datakeskuksella Luulajassa on jo aurinkopaneelit käytössä, ja nyt tutkitaan tuulimyllyn lisäämisen vaikutusta mikroverkkoon. Tässä työssä kehitettiin koneoppimismalli tuulivoiman tuotannon ennustamiseksi datakeskusta varten.

Datakeskuksella on useita sovelluksia tuulienergian ennustamisen hyödyntämiseksi. Uusiutuvat energialähteet ovat luonteeltaan vaihtelevia, joten tuotetun tehon tarkka ennustaminen vähentää ylimääräisen säätämisen ja reservitehon tarvetta sähköverkossa yleensäkin. Datakeskus voi varata uusiutuvaa energiaa markkinoilta seuraavaksi tunniksi tai päiväksi uusiutuvan energian käytön maksimoimiseksi. Ennustaminen 30 minuutista 6 tuntiin etukäteen mahdollistaa työjonon aikatauluttamisen uusiutuvien käytön optimoimiseksi ja vähentää tehonkulutusta. Datakeskus voi pyrkiä minimoimaan sähkön käytön kustannuksia, tai pienentämään kasvihuonekaasujen päästöjä käyttämällä mahdollisimman paljon uusiutuvaa energiaa. Tekoälyyn perustuva älykäs mikroverkko on tapa toteuttaa edellä mainitut sovellukset.

Tutkimuksessa on käytetty kahta avointa tietoainestoa Intiasta ja Ruotsista. Saatavilla oleva data tukee tilastollisen ennustemallin valintaa. Tässä työssä käytettiin satunnaismetsämenetelmää. Intian dataa käytettiin mallin kehityksessä yhtä tuulimyllyä varten. Kehitetty malli ennusti tuotetun tehon hyvin. Ruotsalainen data perustuu EEM20-kilpailuun, jossa arvioitiin koko Ruotsin tuulivoiman tuotantoa. Sitä olikin sovellettava Luulajassa olevan yhden tuulimyllyn tuotannon arvioimiseksi. Luulajan hinta-alueen tuottama teho keskiarvoistettiin, ja ennustamista varten valittiin maantieteellinen paikka läheltä Luulajaa. Kuten oli odotettavissa, soveltamisessa tehdyt likiarvoistukset pienensivät ennustamisen tarkkuutta, jota voidaan kuitenkin pitää kohtuullisena.

Kehitettyä mallia sovellettiin RISE:n ICE datakeskusta varten. Algoritmin validointi on suoritettu, mutta lopullinen testaus tehdään RISE:n simulointiympäristössä. Yleisesti ennustamiseen soveltuvaa dataa ei ole Pohjois-Ruotsista tarjolla. Tieteellisiä artikkeleita ko. maantieteelliseltä alueelta ei löytynyt kirjallisuustutkimusta tehtäessä. Tutkimus ruotsalaisella datalla toi ymmärrystä siihen, mitkä muuttujat ovat merkittäviä Pohjois-Ruotsin alueella sekä niiden suhteellisesta merkityksestä. Kahden eri maantieteellisen alueen tietoaineiston käyttö osoitti, että ilmastolla on huomattava vaikutus koulutetun mallin suorituskykyyn. Näin onkin mielekästä käyttää koulutettua mallia vain sellaisilla alueilla, joiden sääolosuhteet ovat samankaltaiset.

Avainsanat: tuulienergia, tuulivoima, ennustamismallit, satunnaismetsä, datakeskus, tehonkulutus, uusiutuva energia, kestävä kehitys

TABLE OF CONTENTS

FOREWORD

Last year has been interesting experience with distance working. Starting a new job with new colleagues had its own challenges. Our weekly morning coffee meetings enabled to have social contacts with DataAI team and to get to know the people in the team. We have also had biweekly virtual meetings with ArtiqDC project team. It has been useful to hear what is happening in the project, what others are doing, and to give and receive comments about each other's research. Technical challenges with university's different tools in the beginning got solved quite quickly.

I did not know anything about wind power, forecasting models nor data centers, when I started in the project. There was much to learn about those subjects. Now I can say that I have learned about those areas a lot, although that is never ending story. The more you know, the less you know as it is said. Luckily for me I have soul of an eternal learner.

I wish to thank Biomimetics and intelligent systems group (BISG) and ArtiqDC project about possibility to do the thesis. Jaakko Suutala and Satu Tamminen have given guidance and valuable comments to do the thesis. Co-operation with RISE connected the research with data centers more deeply. Mattias Vesterlund and Mikko Siltala have been key persons about issues concerning data centers.

Oulu, 27th of May 2021

Riitta Rankinen

ABBREVIATIONS

AI	artificial intelligence
ArctiqDC	Arctic data centers project
AWPPS	Armines wind power prediction system
CPU	central processing unit
DVFS	dynamic voltage and frequency system
EEM20	17th International conference on the European energy market. 16 – 18 September 2020, Stockholm, Sweden
EM	energy management
FMI	Finnish meteorological institute
GHG	greenhouse gas
HIRLAM	High resolution limited area model
ICE	Infrastructure and cloud research & test environment. ICE data center
ML	machine learning
NetCDF	network common data form
NN	neural network
NREL	National renewable energy laboratory in USA
NWP	Numerical weather prediction
PM	power management
REC	renewable energy credit
RFR	random forest regression
RISE	Research institutes of Sweden
SE1	Electricity price region 1 in Sweden, Luleå
SE2	Electricity price region 2 in Sweden, Sundsvall
SE3	Electricity price region 3 in Sweden, Stockholm
SE4	Electricity price region 4 in Sweden, Malmö
SMHI	Swedish meteorological and hydrological institute
SOWIE	Simulation model for the operational forecast of wind energy production
TSO	transmission system operator
VM	virtual machine
VPP	virtual power plant
VTT	Technical research center of Finland (Valtion teknillinen tutkimuskeskus)
WPPT	Wind power prediction tool

1. INTRODUCTION

1.1. Motivation and context

Rapid growth of internet, cloud-based computing and social media requires increasing amount of data processing power [1]. Number and size of data centers has been proliferating to meet the demand. Figure 1 shows the growth of the cloud data center traffic [2]. To support all the traffic data centers need a lot of energy to function. Year 2019 their consumption was 200 TWh, which is about 1% of all electricity consumption globally [1, 3].

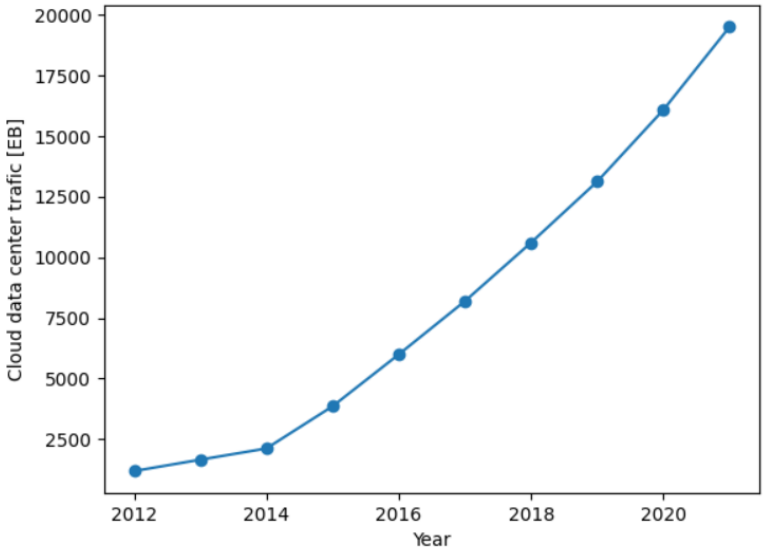


Figure 1. Growth of cloud data center yearly traffic.

New data driven business models have made data one of the key assets in business. As a consequence, to be economically and operationally feasible data centers are increasingly forced to use artificial intelligence (AI) and machine learning (ML) in operations. AI is autonomously taking care of many tasks, such as server optimization and equipment monitoring. Google, for example, is using artificial intelligence to improve efficiency of data centers: cooling of servers and maximum use of renewable energy [4].

Corporate sustainability and social responsibility, including need to reduce carbon emissions, have forced technology companies to invest on renewable and clean energy to run data centers. Companies like Google [5, 6], Apple [7], Facebook [8], and Amazon [9, 10] have been leaders in striving for carbon neutrality. Renewable energy sources are an essential element in achieving the goal. For example, Google is committed to match all energy consumption - in data centers or other facilities – by purchasing equivalent amount of energy from renewable sources, such as wind and solar [5, 6].

Wind power is one important way to produce renewable energy. During year 2020 production in Finland was 7.8 TWh, which is about 10% of all electricity consumption [11]. Production is expected to continue the growth. Forecast for year 2030 is up to 30 TWh for Finland [12]. Figure 2 shows realized and forecasted wind power production

in Nordic countries [12] and Figure 3 realized and forecasted wind power capacity in Finland [13].

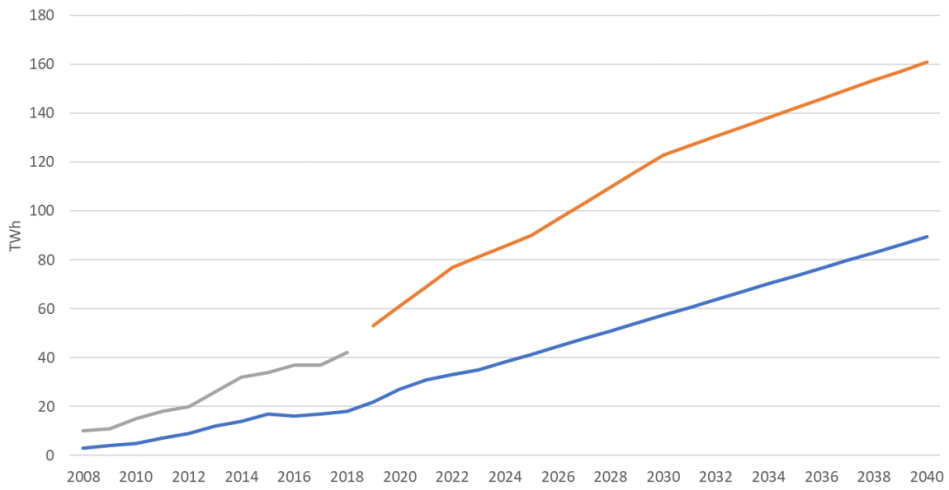


Figure 2. Wnd power production in Nordic countries (gray historical, orange forecasted) and in Sweden (blue) [13].

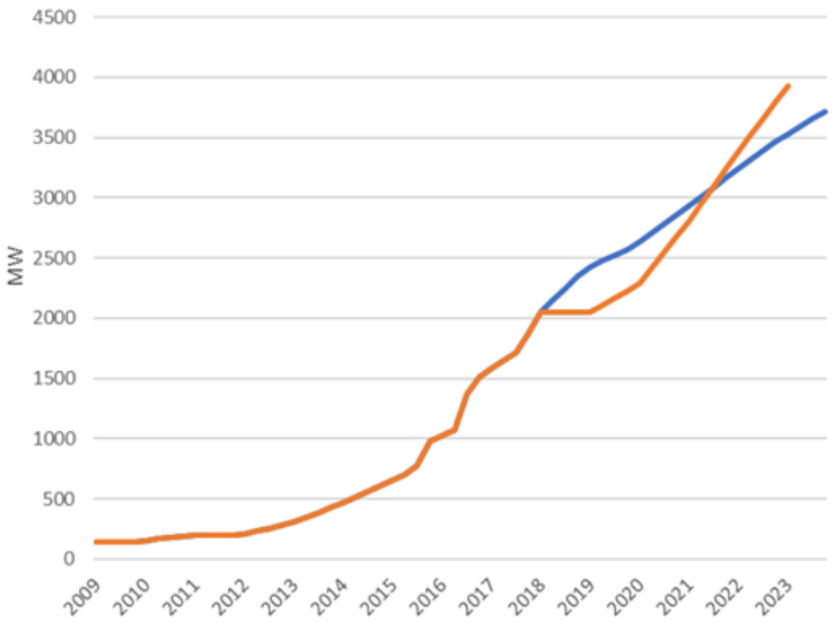


Figure 3. Wind power capacity in Finland: Finnish wind power association based on built and ongoing projects (orange line), and Refinitiv forecast (blue line) [13].

The thesis is part of Arctic datacenter project (ArctiqDC). Co-operation between Nordic regions of Finland and Sweden, and to develop know-how in technical and economical solutions for data centers in arctic area are main goals. Reliable and renewable energy sources are among key factors when Nordic countries are competing on data center investments globally [14]. Especially, hydro and wind power are the main renewable sources in Nordic region.

1.2. Structure and contribution

In the thesis, a machine learning model and algorithm to forecast wind energy for a data center have been developed. Data from Northern Sweden is not openly available. In addition, any scientific articles from the area were not found. The study with Swedish data gave understanding, which variables are significant in northern Sweden and about their relative relationships. Indian data gave possibility to develop the model for one wind turbine. It also proved that climate has a major impact on the trained model, and in which kind of locations it is reasonable to use for forecasting.

Chapter 2 covers renewable energy sources and how to produce energy with them. Main focus is on wind energy. General structure of PM/EM system of a data center is presented. Smart microgrids, which control PM/EM system having renewable energy sources, are discussed.

Chapter 3 explains main methods used to forecast wind energy production. First numerical weather model is shortly explained (NWP). NWP forecasts are used in wind energy forecasting. Physical, machine learning and combined methods used in wind energy forecasting are presented.

Chapter 4 presents and analyzes data sets used in the thesis. First set is data of a wind turbine in India. It is used to develop prediction mode especially for one wind turbine. Second data set is EEM20 competition data covering weather of Sweden, years 2000 and 2001. Total production data of Swedish price regions is used to estimate wind energy production of one wind turbine in Sweden. Set is important since wind energy production is heavily dependent on local weather.

Chapter 5 covers development of models for both Indian data and Swedish competition data. Model development includes defining hyperparameters, checking matching for overfitting and feature importances. Feature sets are selected for wind power forecasting experiments, which are then computed.

Chapter 6 describes how developed model is applied to RISE's ICE data center. Structure of facilities of the data center is shortly described. Then adaptation of local weather forecasting parameters for testing are presented. Code used to train and test the model is explained. Results of forecasting experiment are shown. Importance of sustainability and renewable energy sources related to a data center is noted. RISE's plans about usage of the developed model is explained. Finally, discussion of research is in Chapter 7 and conclusions in Chapter 8.

2. SUSTAINABLE ENERGY

2.1. Renewable energy

2.1.1. Wind energy

Wind power forecasting has many both technical and economical applications. Tables 1 and 2 show major ones, and their time horizons as defined by Heinermann [15] and Finland's technical research center (VTT) [16], respectively.

Trading at electricity market has important use cases. Energy is sold at future point of time, so forecasting is required to estimate the future price. Time horizon of forecasting ranges from a few seconds up to a day depending on type of trading (Table 1). For example, European power exchange Nord Pool offers clearing, intraday trading and one day-ahead trading [17].

Renewable energy sources are intermittent. Accurate wind power forecasting reduces the need for additional balancing energy and reserve power to integrate wind power. Balancing electricity grid means that frequency has to stay e.g. in range of 49.9 Hz – 50.1 Hz. To achieve this transmission system operators (TSOs) such as Fingrid in Finland is using energy bids and energy prices [18], i.e. operator has to manage demand and supply. Battery packs are used to balance electricity grid as well. For example, Tuuliwatti has built 6 MW battery pack next to Viinamäki wind farm in Ii, Finland [19]. French company Neoen is building 30 MW battery pack in Lappeenranta, which will co-operate to balance Fingrid's electricity grid [20].

Table 1. Horizons in forecasting wind power defined by Heinermann [15]

Horizon	**Time range**	**Applications**
very short-term	few seconds – 30 min	market clearing trading balancing virtual power plants
short-term	30 min – 6 h	load balancing intraday trading regulation
medium-term	6 h – 1 day	day-ahead trading price optimization
long-term	1 day – ≥ 1 week	planning of reserve energy scheduling of maintenance

Table 2. Horizons in forecasting wind power defined by VTT [16]

Horizon	**Time range**	**Applications**
very short-term	minutes – hours	controlling wind turbines
short-term	up to 72 hours	used by TSOs and energy traders
medium-term	1 week	scheduling maintainance plans

Wind farm operators need to maintain production equipment. Very short-term forecasting is used to control turbines. Long-term forecasting allows to schedule maintenance breaks to those time points where production is forecasted to be low.

Virtual power plant (VPP) is a distributed power plant, which collects different medium or small size power plants to function as one plant. Separate plants are still independent in their operation and ownership. Wind farms, solar parks, combined heat and power units, flexible power consumers and storage systems may be part of VPP. Forecasting wind energy is necessary in managing VPPs whenever wind farms are included [21, 15].

Wind energy is produced by wind turbines, which are usually located on wind farms each having up to hundreds of turbines. Wind farms can be onshore or offshore. In Finland, the farms are mostly onshore because of economic reasons. Sweden on the other hand is economically supporting cabling of offshore wind farms.

Wind turbines are built on various sizes and technologies. Most common type has horizontal axes with 3 blades. Wind turbines have been getting bigger over years. Diameter of a rotor can be over 150 meters and power 5 MW onshore and more than 10 MW offshore. Height of some latest wind turbines can even be over 200 meters, although around 180 meters is more typical. The "engine" turns the turbine towards wind to maximize efficiency.

Turbine is transforming wind speed into electric power. Typical power curve is shown in Figure 4. Cut-in speed is typically 3 – 4 m/s. Then power increases following mild S shape until maximum speed is reached, typically 11 – 17 m/s. If wind speed exceeds cut-out speed, typically 25 m/s, rotor is switched off for security reasons. Naturally, wind farms do not operate on maximum power all the time. Typical utility level in Finland is 24% – 40% [22].

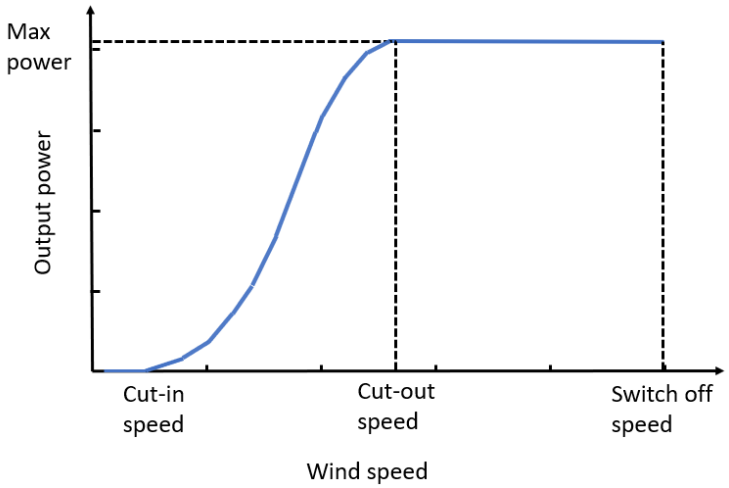


Figure 4. Typical power curve of a turbine.

In Finland, there is more wind in winter than summer. Additionally, air in winter is denser than in summer. As a result, 70% of production is produced during 6 of the coldest months [23, 24]. During winter time, blades may get icing, and even small amount of ice can impact on produced power level. One solution is to have a heating system installed on the blades.

2.1.2. Other renewable energy sources

Solar energy capacity has high growth rate in Finland. At the end of year 2019, small scale capacity connected to electricity network was 198 MW, where increase was 77 MW (64%) in one year [25].

Solar energy is typically produced using photovoltaic panels. Panels have low operating and maintenance costs. There are no moving parts, and they can be used 20 – 30 years. About 21% of radiation can be transformed to electricity. Weather dependence is natural, either sun is shining or not, and cloudiness and radiation level vary over time. The best season for solar power is from March to October, when summer days and nights have a lot of light. Drawback in Nordic countries is winter time with less daylight. In addition to roofs, panels can be installed on facades as well. In winter, sun is shining from lower angle, so panes on facades are able to collect more sunlight than panels on roofs.

Hydro power has been used in Nordic countries already a long time. Capacity in Finland is about 3190 MW. Hydro power covers about 10% - 15% of electricity production in Finland. Most of rivers have already been harnessed for power production, so there is no room to increase production any further. Hydro power is quite steady energy source compared to wind and solar power, so balancing of electricity grid is easier.

Other renewable energy sources such as marine (tidal, waves) and geothermal energy, are not used in Finland to produce electricity.

2.2. Data centers

A data center is a physical facility to house applications and data of organizations. It has a network of computing and storage resources, which enable to share applications and data. The main components are routers, switches, firewalls, storage systems, servers, and controllers for application-delivery.

2.2.1. Power consumption and power management

Sizes of data centers range from a few kW to more than 100 MW. They consume a lot of power and thus produce a lot of heat. Most of the power is used by IT equipment (40% - 50%), and cooling (about 40%) [26]. Values depend a lot on type and size of the data center. See Figure 5 for a typical breakdown.

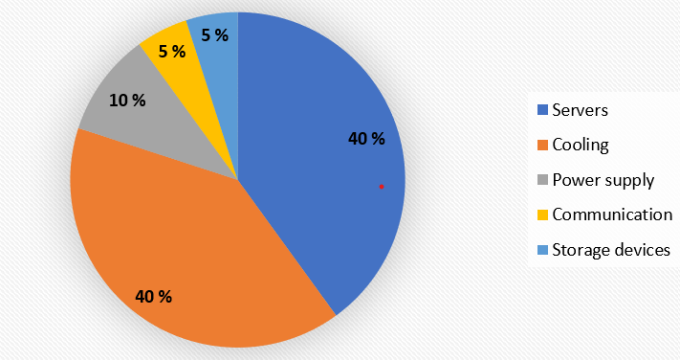


Figure 5. Power usage breakdown of a data center

Power consumption can be controlled in many ways. Traditionally, servers are switched on and off depending on the data load. In order to do this, jobs are scheduled [27]. Powered-on idle server consumes typically half of its peak power consumption.

Hardware of servers can be optimized for lower consumption as well. For example, in addition to setting processors to deep sleep or deeper sleep, their consumption can be controlled by dynamic voltage and frequency system (DVFS). Placement and consolidation of virtual machines (VM) affects directly to number of powered-on servers [27]. VM is a collection of cloud infrastructure resources that are specifically designed for business needs. Resources here cover CPU, memory, disk space and bandwidth.

One key issue of any data center is resiliency, i.e. data center has to be able to maintain required service level during any disruptions [28]. Outages are very expensive - average outage cost 2015 was $740,357 [28]. Thus, data center needs a back-up system to stay up. Figure 6 shows typical power management system. During outage microgrid controller isolates data center from main grid, and energy is commonly used either from battery pack or diesel generators. Only one equipment is used at a time. Another method for resiliency is to double some or even all of the equipment, but naturally this is very costly.

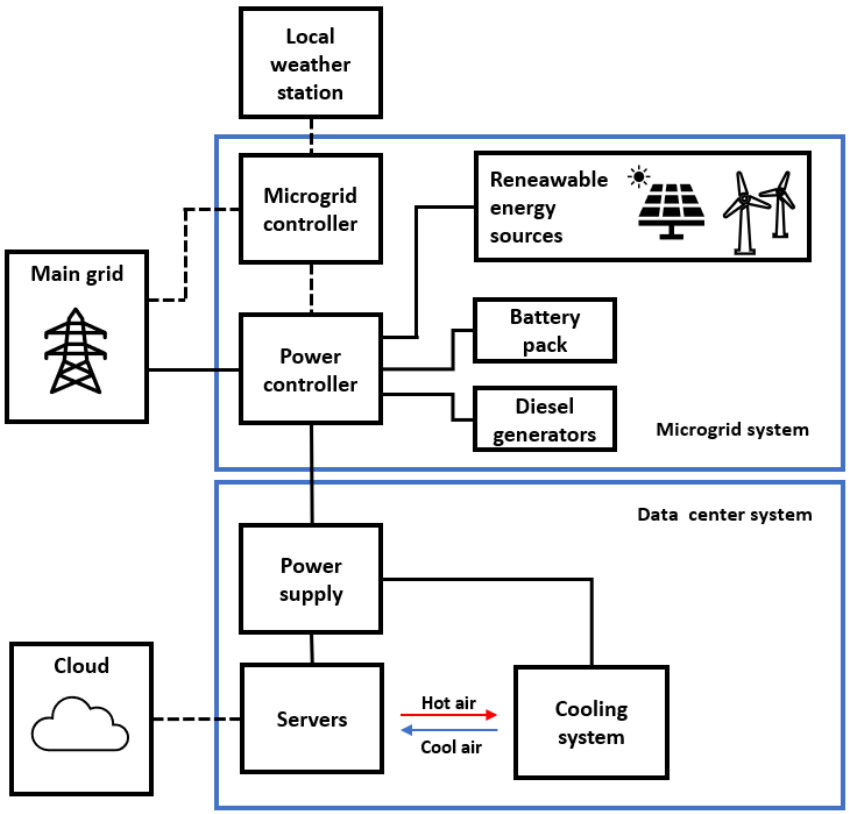


Figure 6. Power management system of a data center.
Dashed line = data, solid line = electricity

2.2.2. Smart microgrid

Sustainability trend has increased the usage of renewable energy sources significantly. The easiest way for data center companies to reach for carbon neutrality is to purchase renewable energy credits (REC) to compensate emissions from diesel generators and energy from main grid, which is often generated by fossil fuels. To

actually use renewable energy company has to buy it from elsewhere or to produce it in its own microgrid.

Smart microgrid is needed for several purposes. Naturally, it is managing data center's grid. Microgrid takes care of balancing including job scheduling and taking care of variable power from renewables. In case of any disturbances, such as power outage, microgrid can isolate into an island mode, and then use own renewable energy sources or energy from battery pack.

Smart microgrid can co-operate together with main grid. It enables participating demand & response programs, and thus selling or buying energy from main grid. Battery pack is able to store excess own renewable energy for later use or trading. Real time changes of power prices can be tracked to avoid peak times on the main grid for lower cost.

Data center's operations can be optimized to minimize energy costs, or to minimize emissions of green hours gases. Maximizing use of renewable energy, either own or bought, helps to minimize GHG emissions. One way to maximize is opportunistic job scheduling where renewable energy is used whenever it is available [27].

Core of a smart microgrid is advanced software and controls. All functions are basically automatized. Artificial intelligence, i.e. data driven solutions, data analysis and machine learning are used in algorithms to realize forecasting and optimization.

2.2.3. Artificial intelligence

Many AI and ML implementations in data centers are already in use and there are more advanced ones in development. Applications covered here are: reducing power consumption, reducing downtime, optimizing servers, monitoring equipment, and security issues [29].

AI is able to learn temperature set points, test flow rates and evaluate cooling equipment. It can also be trained to collect critical data from sensors. Sources of energy inefficiency can be found, which allows AI to autonomously reduce energy consumption. Cooling system is a major consumer of power in any data center. Google was able to reduce energy consumption of their cooling system by 40% with AI.

As already discussed, outages are very expensive. AI can monitor server performance, network congestions, and disk utilization to detect and predict outages. Predictive analysis can track power levels and identify potential problem areas. For example, AI can predict and identify outages and recognize users that might be affected. This information allows AI to implement autonomously mitigation strategies for easier recovery.

AI can help to optimize operations of servers. Job scheduling means that workload is distributed across various servers supported by predictive analysis. Load balancing algorithms learn from past data to do this effectively. Optimization can speed up finding flaws, reduce processing times, and resolve risk factors.

Equipment failures are expensive, since failed equipment need to be repaired or replaced, or otherwise they could cause downtime. Heat produced in a busy data center makes a risk of failures higher. Flaws in cooling system may lead to overheating and shutdown. AI can identify defects in equipment by using pattern-based learning. Smart sensors are installed to equipment for monitoring for any abnormal vibrations or sounds, which enable AI to predict potential equipment failures.

Data in data centers is usually confidential, some of it even business critical. It needs to be protected from unauthorized users. AI can learn from normal network behavior

and detect cyber threats based on deviations from normal behavior. AI can also detect malware and identify any security flaws in data center systems. Incoming and outgoing data can be screened thoroughly for security to detect any threats.

2.2.4. Wind energy

Wind energy forecasting has several applications in data centers. Some examples are presented here. Renewable energy from market can be reserved for next hour or next day to maximize its use [6]. Forecasting from 30 min to 6 hours ahead allows job scheduling to optimize a usage of renewable energy. Both trading energy to and from main grid and maximizing a use of renewables support a goal of carbon neutrality, or on the other hand, to minimize energy cost.

Since in Nordic countries, there is more wind energy production in winter [23, 24], and more solar energy production in summer due to longer daylight [30], combining both energy sources into the same microgrid may even up seasonally changes of power production in electricity grid.

3. FORECASTING METHODS

Wind power forecasting models can be divided into three main groups: physical models, statistical models, and combined models. The models are using weather forecasts, which are provided by meteorological institutes or other weather services. Quality of weather service has a major impact on wind power forecasting.

3.1. Physical models

Numerical weather prediction (NWP) is based on mathematical models. Computations following physical laws calculate the state of the atmosphere. Navier - Stokes equations describe motion of viscous fluids such as liquids and gases. Computations on complex models are heavy and made with supercomputers [16]. Weather predictions are done globally and locally. There are many different NWP models used around the world. Finnish and Swedish meteorological institutes use High resolution limited area model (HIRLAM) in regional forecasting [31].

NWP forecasting is typically using ensemble NWPs. By averaging ensembles forecasting gains higher accuracy. There are two ways to obtain ensemble NWPs: running differently calibrated NWP models or varying slightly the initial conditions.

Physical models use parameters of local terrain and wind farm [32, 16]. Model to forecast wind power production is shown in Figure 7. Starting point is NWP weather forecast, which is delivered in resolution defined by calculation grid. Wind farm description (layout, etc.) and terrain description (orography, roughness, obstacles, etc.) are used to transform weather parameters into local wind speed. Hub height and power curve of a turbine are then used to calculate produced output power. In case on-line data is available, statistics can be calculated to reduce error of forecast [32, 16].

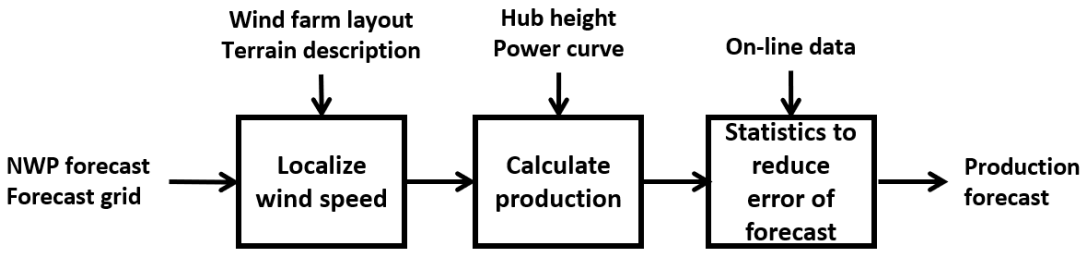


Figure 7. Physical model to forecast wind power production.

Accuracy of physical models depends on a weather data available. Influence of athmospheric dynamics becomes more important in very short-time and short-time horizon. This means that if weather data is available in short intervals, then also prediction of a physical model is better [32, 16]. Open weather data is delivered typically on 10 minute intervals.

Purely physical models are not so popular anymore. Many of them have been developed further to be part of a combined method. For example, an older model Prediktor is nowadays part of Zephyr [33, 34]. Previento has similar principle than Prediktor. Previento uses neural networks (NN), which takes on-line wind generation measurements as inputs. Previento has regional forecasting and uncertainty information, which are missing from Prediktor. Simulation model for the operational forecast of wind energy production (SOWIE) is a commercial solution today [33].

Windpowerlib is an open source python library for generating wind feed-in time series. Weather data, wind turbine and wind farm parameters are fed in to calculate power output [34]. Table 3 shows the physical methods.

Table 3. Physical forecasting models

Model	Developer	References and comments
Prediktor	Risø, Denmark	Now part of hybrid model Zephyr [33] [34]
Previento	Oldenburg University, Germany	[35] [77]
SOWIE	Eurowind GmbH, Germany	[33] [34]
windpowerlib	open source library	python library [36]

3.2. Statistical models

Statistical relationship is developed between weather forecast and power output for a wind turbine or wind farm. A lot of historical weather and power output data is used to train the model. Typical development and usage process for a machine learning model is shown in Figure 8. To be effective statistical learning requires pre-processing of input data, so called feature representation. Data needs to be "cleaned" from missing or false measurements, outliers, etc. Features for the model are then extracted from "clean" data. Features should represent the data well, so that forecasting will be reliable. Data is split for training and validation. Then chosen model is trained and its hyperparameters are tuned. After testing, model is ready to be used in forecasting. In wind power case, NWP data is fed to tuned model for forecasted output power.

Processing is usually time consuming and requires good understanding of the subject area and use of right algorithm to optimize data transformation. Exception is neural networks (NN), which do the feature representation automatically. NNs are so called "black-box" models, since user does not know what is going on inside NN.

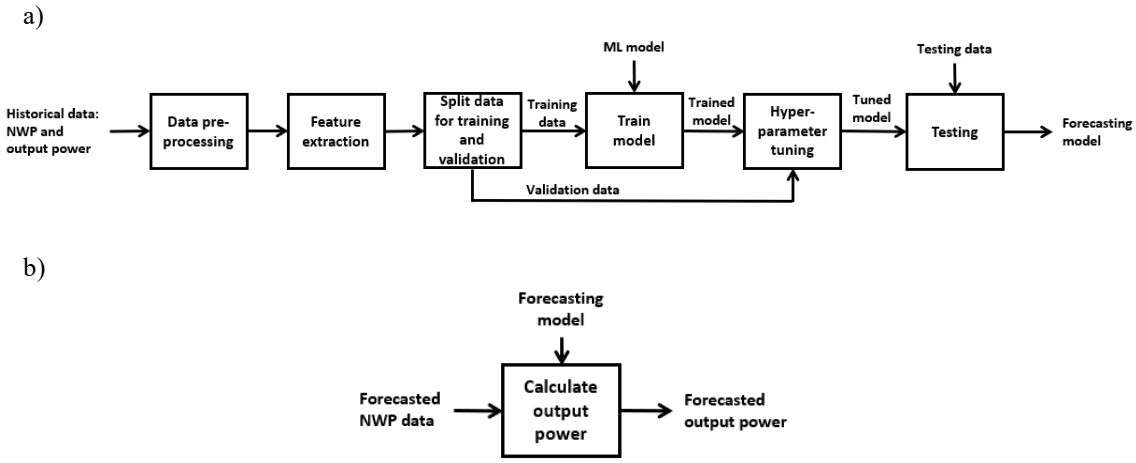


Figure 8. a) Development of a machine learning model. b) Usage of trained and tuned forecasting model.

3.2.1. Conventional statistical models

Conventional statistical models are based on classical linear statistical models: auto-regressive (AR), moving average (MA), auto-regressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and seasonally adjusted ARIMA (SARIMA) [32]. Classical models are easy to formulate, and they produce timely forecasts. However, they look backwards, so they are inherently good in predicting future in case there is a steady trend, not if there is a turning point. Like with all statistical models, model parameters have impact on forecasting accuracy, so they need to be selected carefully.

Several ways to improve forecasting accuracy have been studied. Milligan et al. [37] have used ARMA model to forecast wind speed and output power in a wind farm in USA. Performance was highly dependent on chosen parameters. Model by Liu et al. [38] is using wavelet decomposition together with ARIMA. First wind speed is forecasted from sub-series from wavelet decomposition. Then final forecast of wind speed is done with aggregate calculation, which improves the accuracy compared to the classical models. Autoregressive conditional heteroscedastic model combined with ARIMA (ARIMA-ARCH) uses wavelet decomposition as well to forecast wind speed [39]. Also, Bayesian methods have been studied. Miranda et al. [40] used AR model based on Bayesian approach to forecast wind speed 1 hour ahead. Combined model of Bayesian clustering by dynamics (BCD) and support vector regression (SVR) by Fan et al. forecasted wind power in short-term for a wind farm [41]. Wang et al. [42] have combined extreme learning machine (ELM) algorithm with Ljung-Box Q test (LBQ) and SARIMA model. Accuracy was improved compared to single classical models.

Classical models are mostly used for very short-time and short-time forecasting. Main applications for these time ranges are presented in Section 2.1.1 (Tables 1 and 2). Forecasting a wind speed is much more common than forecasting a produced output power. Models are popular as reference models, especially the ones with improved accuracy. As presented, classical statistical models are in many studies also combined with machine learning techniques.

3.2.2. Random forest regression

Random forest regression (RFR) is widely used algorithm for regression problems in general. RFR is considered to be one of the most accurate learning algorithms. It is efficient with large data sets, and also able to handle even thousands of input variables. Like other decision trees RFR can calculate feature importances, so that selection of feature set is easier [43]. RFR can effectively handle large amounts of missing data, and still maintain accuracy. Disadvantages are danger of overfitting, and biasing in case of including categorical variables.

Method uses voting mechanism by calculating average of predictions (Figure 9). First n decision trees (i.e., estimators) are built with specified hyperparameters. Each decision tree predicts a number as an output for a given input. Predictions of the estimators are then averaged. Averaging enables higher accuracy [44].

Decision trees overfit very easily so stopping criteria is necessary. One common way is to define maximum depth, which allows only a certain number of splits from the root node to the terminal nodes. Another way is to set a minimum number of samples in each terminal node to prevent splitting going too far.

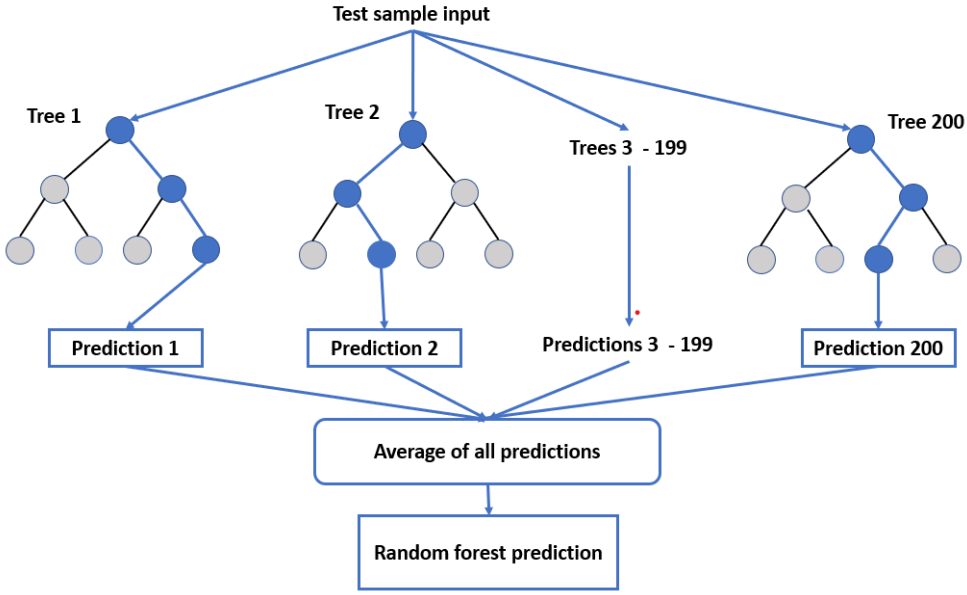


Figure 9. Random forest regression method. Number of estimators 200.

Lahouar et al. [45] have used RFR to forecast wind power one hour ahead. Their prediction is more accurate compared to classical neural network. Wind speed and wind direction are used as features in forecasting. Short-term prediction model by Zhou et al. [46] is based on RFR as well. Achieved accuracy was better compared to classical models. Fischer et al. [47] have used CART-Bagging algorithm similar to RFR. CART (classification and regression tree) is a decision tree algorithm, which is bagged (bootstrap aggregation) for better accuracy. This work applies RFR in experiments in Chapter 5.

3.2.3. Support vector regression

Recently, there has been a lot of research on support vector regression (SVR) for wind power forecasting. SVR is easy to implement and insensitive to outliers. Decision model is easily updated. Prediction accuracy can be improved by measuring the confidence. Drawback is that computer load of SVR is higher compared to other regression models [15]. Research has been mostly about improving performance of basic SVR by combining it with some other methods.

Heinermann [15] has studied RFR, SVR and k-Nearest-Neighbors (k-NN) models for wind power forecasting. RFR and SVR offered good prediction accuracy, while k-NN proved to have clearly poorer performance. He used SVR ensembles to improve quality of forecasting and reduce computation time. SVR ensemble method using bagging and weighted averaging had better accuracy and reasonable runtime compared to SVR. Combination of SVR and decision tree gave 37% improvement in accuracy and some in computer time.

Kramer et al. [48, 49] have predicted wind energy using SVR method as such for short-term forecasting. Only wind measurements from wind turbines were used, no meteorological data. Model predicted well. Kernel method was combined with SVR non-parametric density estimation to help with the dynamics of wind time-series data

[50]. First wind data was modeled with kernel density estimation, and then wind energy predicted with SVR.

Zeng & Qiao's method had two steps [51]. First wind speed was predicted with SVR. Then wind power was predicted by using characteristics of the wind turbine. Botha & van der Valt [52] have showed that careful selection of features improves performance of SVR by 11.2%. Li et al. [53] combined SVR with Dragonfly optimization algorithm [54]. Prediction performance was better compared to e.g. standard NN trained with backpropagation or Gaussian process regression.

Zameer et al. [55] proposed machine leaning based short term wind power prediction model (ML-STWP). Method is hybrid model using feature selection through irrelevancy and redundancy filters. SVR is then used for auxiliary prediction. Wind power is predicted with enhanced particle swarm optimization and hybrid NN.

3.2.4. Neural networks

Neural networks (NN) have been widely researched for wind power forecasting during recent years. NN is able to model a complex non-linear relationship. Training defines dependence between input and output, implementation is simple, and it is fast to develop. Exact mathematical equations are not necessary. Accuracy of NN solutions has been improved in many ways.

Manero Font [56] has compared different NN architectures. He proved that separable CNN is the best architecture for wind power forecasting. Other architectures have also shown excellent performance, but with some added computing costs. RNN LSTM and RNN GRU are popular structures in time-series prediction. However, they were not so successful in wind power forecasting [56].

Hybrid approach of principal component analysis (PCA) and deep learning was developed by Khan et al. [57]. PCA extracts hidden patterns from wind data, identifies meaningful information, and removes high correlations between values. Neural network and algorithm predicts the wind signals. Accuracy of the model was better compared to standard NN trained with backpropagation, SVM, ensemble selection, CNN, and AR. Ghadi et al.'s [58] model is using evolutionary algorithm to optimize training of NN. Accuracy of prediction for short-term and very short-term is enhanced with the method.

Grassi & Vecchio [59] designed a two hidden layer NN in combination with back propagation learning algorithm. Architecture is using hyperbolic tangent transfer functions in first hidden layer and logarithmic sigmoid transfer function in the second hidden layer. Predicted energy values were in line with measured values. Diaz-Vico et al. [60] managed to improve accuracy by using NN ensembles. Wang et al. [61] used NN based ensemble approach, where wavelet transform decomposes raw data into different frequencies. Nonlinear features of each frequency are then learned by convolutional NN to improve accuracy.

Kitajima et al. [62] used complex-valued NN to predict wind power. Instead of wind speed and direction wind was treated as a complex number. Compared to real value complex-valued NN prediction was more accurate. Liu et al. [63] developed a two-step process. First probabilistic NN filters raw data, which had been collected by plant's information system. Valid data from step one is then used in step two, where model to predict wind power is built using complex-valued recurrent neural network.

Bilal et al. [64] used neural network for wind power forecasting in Senegal. Their main conclusion was that performance of the model depends on site. Difference is

significantly related to site characteristics and operating mode of the wind turbine. Also, number and selection of features used in the model have significant impact on performance. Bilal's findings are in line with conclusions in the thesis here.

NN and fuzzy logic approach is applied to cases where system is difficult to model accurately, but an inexact model is available [32]. Approximate values and incomplete or ambiguous data can be used. Fuzzy logic alone has weak learning ability though, so it is typically used together with NNs. Research in this area has been done, e.g. by Negnevitsky et al. [65], Sideratos et al. [66] and Damousis et al. [67].

3.2.5. Other statistical models

Many other statistical models have been used in wind power forecasting as well. However, their significance seems to be minor. k-NN model has been combined to other models, since alone k-NN has a poor accuracy [15]. It has been used more as a reference model in research. Jursa & Rohrig [68], Azeem et al. [69] and Zhang et al. [70] have combined nearest-neighbor with NN to achieve better accuracy. Gaussian-process (GP) model based on kernel machines and Bayesian estimation forecasts upper bound, lower bound and average of wind speed [71]. More advanced GP model, called sparse heteroscedastic GP, reduces computational cost [72]. Yan et al. [73] forecasted wind power with temporally local Gaussian process, which examined forecasting errors. Less measurement data was needed and predictions were faster and more accurate than with standard GP.

3.2.6. Statistical models for wind farms and regions

Table 4 presents statistical forecasting models. Many of the methods listed are already commercialized. Armines wind power prediction system (AWPPS) provides short-term forecasts for both off-shore and on-shore wind farms. Short-term forecasts are based on statistical time-series approach. Long-term forecasts are based on adaptive fuzzy neural networks. Short-term and long-term forecasts can be joined by weighing method. System includes uncertainty estimation of confidence intervals based on adapted resampling.

AleaSofts' and University of Catalunya's AleaWind (former GH forecaster) can provide national, regional, or single wind farm forecasts. Model uses NWP data, wind speed and wind direction as input. In addition, on-line and off-line data from wind farm are used. Parameters of neural network with SARIMA structure are estimated on-line, which allows the model to respond continuously to any changes in the system.

DNV provides services in wind power forecasting. Method uses multi-parameter statistical regression to transform global NWPs and site data into site-specific models. The models can perform user-defined transformations between NWP and the site.

Sipreolico is using several adaptive statistical models. Spanish NWP forecasts and hourly SCADA data from all Spanish wind turbine are fed to adaptive statistical models that produce a final forecast using an adaptive combination of the alternative predictions. Method can adapt to changes in the operation of wind farms or in the NWP prediction model. No precalibration is needed.

Wind power prediction tool (WPPT) is using reference wind farms in forecasts. Inputs for the model are NWP forecasts for the region and reference farms, and on-line data from reference farms. Further inputs are max power and operation times of

farms, and aggregated energy readings from turbines. WPPT has two branches for regional forecast. First forecast is made for each reference farm, and then upscaled to cover sub-area and further whole region. Second forecast of each sub-area is computed directly with off-line measurements of sub-area output power and relevant NWP data. Finally for the whole area, output power is weighted with average of forecasts from both branches using RMSE criterion. WPPT provides statistical models for short-term predictions of the wind power production in wind farms or areas. Time-adaptive and recursive estimation method allows system to adapt to changes, such as growth of surrounding vegetation.

Table 4. Statistical forecasting models

Model	**Developer**	**References and comments**
AWPPS	Ecole des Mines de Paris, France	[74] [75] [76] [77]
AleaWind	AleaSoft, Universitat polytechnica de Catalunya	[78] [77]
DNV forecaster	DNV GL, Norway	[79] [77] former GH Forecaster
Sipreolico	University Carlos III & Red Electrica de Espana	[80] [77]
WPPT	Eltra/Elsam & IMM/DTU, Denmark	[81]

3.3. Combined models

Many models have been combined to take advantages of each individual model in the combination. Target is to gain optimal performance and to improve prediction accuracy. Combined models are also called hybrid models. There are many types of combinations in wind power forecasting [77]:

- physical and statistical approaches
- short-term and medium-term models
- alternative statistical models
- alterative artificial intelligence models

Table 5 presents forecasting models combining physical and statistical approaches. Many of them are already commercially available. In addition to ones listed below, there are models developed in USA, but there is not much information available about them.

ANEMOS is an EU project of 26 partners. Nine previously developed models are used in the project to develop more accurate model for onshore and offshore forecasting. Both physical and statistical approaches are included. Emphasis in the project is on complex terrain, extreme weather conditions and offshore prediction. Project is studying economic and technical benefits at different levels: national, regional and a wind farm, and time horizons from minutes to days ahead.

EPREV has a chain of models. Mesoscale models forecast wind conditions 72 hours ahead. Surroundings (topography, roughness, and obstacles) of each wind farm is taken into account. The forecasts are then converted to power using either physical or statistical models. Physical model generates wind farm power curve. Model is based on simulation of atmospheric flow in the wind farm area, and further for each turbine. With help of turbine's power curve and thrust curve, output of the wind farm can be estimated. In addition, wind farm power curve and calculated power series can be used to train statistical models as well. Statistical model uses AR for very-short time forecasting and NNs for short term forecasting. Each turbine is modeled individually.

Truewind's model eWind is running a mesoscale weather model using boundary conditions from a regional weather model. This allows more physical processes to be captured, and the prediction to be more local. Adaptive statistics is used to reduce systematic errors, either multiple screening linear regression model or a Bayesian neural network.

Chinese INT-WPFS combines physical model based on NWP forecast and statistical model. System provides real-time anemometer information, very short-term power prediction and short-term power prediction.

LocalPred was specifically developed for wind farms on complex terrain. Roughness and topography of wind farm modify forecast of MM5 weather model. Model output statistics (MOS) based on fuzzy logic and self-tuning is used to reduce systematic forecasting errors. Forecasts are then transformed into power with statistical model for each wind direction and air density. Linear ARMA models are used for very short-term forecasting. RegioPred extends prediction of a single wind farm to a regional one. The regional forecast is done by summing forecasts of each wind farm or selected reference wind farms using cluster analysis.

Scirocco used three consecutive steps from physical and statistical processes. Physics and mathematics is applied to part of the steps. First model is MOS after first weather prediction to handle systematic errors of NWP. Second model computes local wind from adjusted weather parameters on surrounding grid points, local orography, and local roughness. Third model, which is MOS as well, uses turbine and farm characteristics. System is able to adapt to local geographical and wind farm characteristics.

WEPROG (MSEPS) has are two main models. Weather prediction system uses ensemble technique. Physical parametrizations are used to vary meteorological processes in NWP models, and 75 different forecasts for each model run are produced. Uncertainty is predicted to improve accuracy. Power prediction system is trained first with historical weather and power data. Forecast is computed for each ensemble member separately. Forecasts are made for wind farm or region.

Wind power management system (WPMS) covers over 95% of all wind power forecasting in Germany. WPMS monitors on line current wind power generation for control and subregions, and secondly forecasts wind power day-ahead and short-term for single wind farms, control area and subregions. Numerical mesoscale atmospheric model transforms NWP data to wind farm location. Then data is fed to NN. Forecasts are computed for each wind farm and then summed up and upscaled by a transformation model. NN also provides wind farm power curve.

Zephyr combines statistical WPPT and physical Prediktor models. Each wind farm has own model for forecasting. If only number, type, and location of wind turbines of the wind farm are available, then Prediktor model is used only NWP data as input. Statistical models of WPPT can be used if all data of the wind farm is available including on-line data.

Table 5. Forecasting models combining physical and statistical approaches

Model	**Developer**	**References and comments**
ANEMOS	26 partners from 7 EU countries	[75] [83] EC project
EPREV	INESC, INEGI & CEsA, Portugal	[84] [77]
eWind	AWS Truewind, USA	[83]
INT-WPFS	Chinese Electric Power Science Istitute, China	[85]
LocalPred & RegioPred	CENER & CIEMET, Spain	[86] [83] [77]
Scirocco	Aeolis Forecasting Services, Netherlands	[77]
WEPROG (MSEPS)	University college Cork, Ireland	[87] [77]
WPMS	ISET, Germany	[88]
Zephyr	Risø & IMM, Denmark	[33] [34] combination of WPPT and Prediktor

4. DATA SETS

Two data sets have been used in the analysis: data from a wind turbine in India [89] and data for EEM20 Wind Power competition [90].

4.1. India data

India data covers measurements of a turbine over 2.5 year period from December 2017 to 31 March 2020. Measurements are recorded at a 10-minute interval. Since a lot of data is missing, the usability of the data set is restricted. One year period with better quality was selected for the analysis. Data is available in CVS format [89].

4.1.1. Input and output variables

Data set contains various weather, turbine, and rotor variables (Table 6). Output power, ambient temperature, wind speed and wind direction are typically used as features in wind power forecasts. Rest of the variables are highly turbine specific. Available data allowed to develop model on how produced output power of one turbine depends on weather data.

Table 6. Variables in Indian turbine data [84]

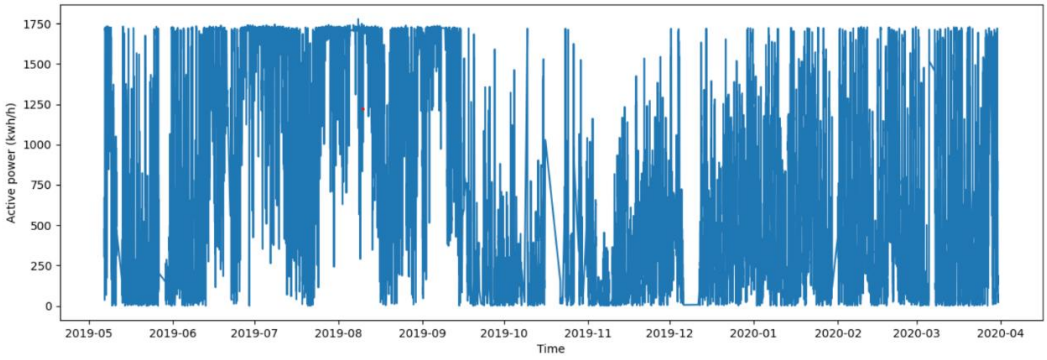
Variables used in thesis	**Variables not used in thesis**
Time Stamp	BearingShaftTemperature
ActivePower [kWh\h]	Blade1PitchAngle
AmbientTemperature [ºC]	Blade2PitchAngle
WindSpeed [km/h]	Blade3PitchAngle
WindDirection [degrees from north]	ControBoxTemperature
	GearboxBearingTemperature
	GearboxOilTemperature
	GeneratorRPM
	GeneratorWinding1Temperature
	GeneratorWinding2Temperature
	HubTemperature
	MainBoxTemperature
	NacellePosition *)
	ReactivePower
	RotorRPM
	TurbineStatus
	WTG Turbine name

*) nacelle = casing for components to turn turbine into the wind

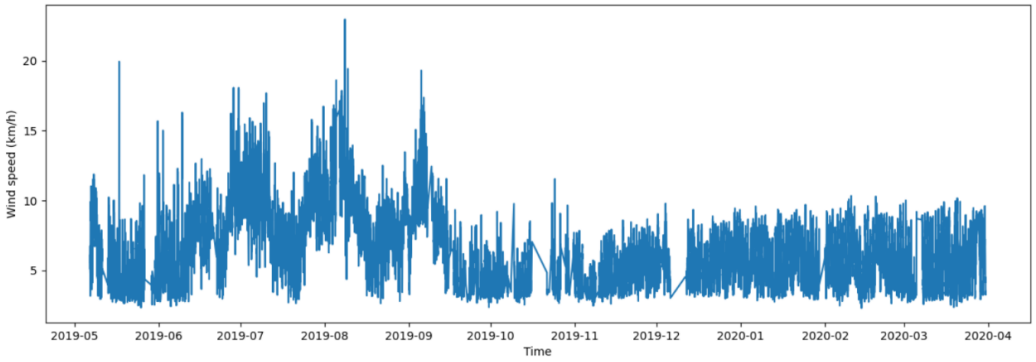
Figures 10 and 11 show values of output power, wind speed, wind direction and air temperature over measurement time. Output power varies from 0 kWh/h up to 1750 kWh/h, which is maximum power the turbine can produce (Figure 10a). Climate in India has seasonal pattern, where wind speed and direction changes. During summertime wind speeds are higher and it blows from north, while during winter wind is calmer and blows from south (Figure 10b and 10c). Wind directions close to 0 and 360 corresponds to north. This makes north wind look chaotic for human eye in the figure (Figure 10c). Thus, zonal and meridional components of the wind were calculated (Figure 11a and 11b). They have similar characteristics with each other.

Temperature range is roughly 22°C – 40°C; cooler in winter and warmer in summer (Figure 11c).

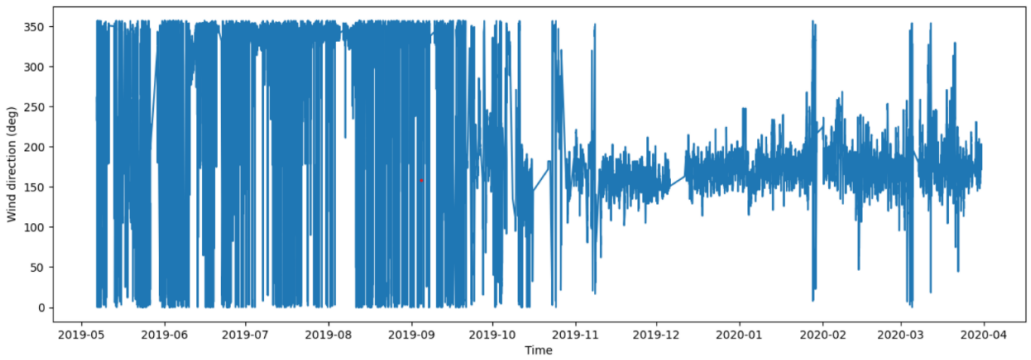
a) Active power



b) Wind speed



c) Wind direction



d) Wind direction, detail

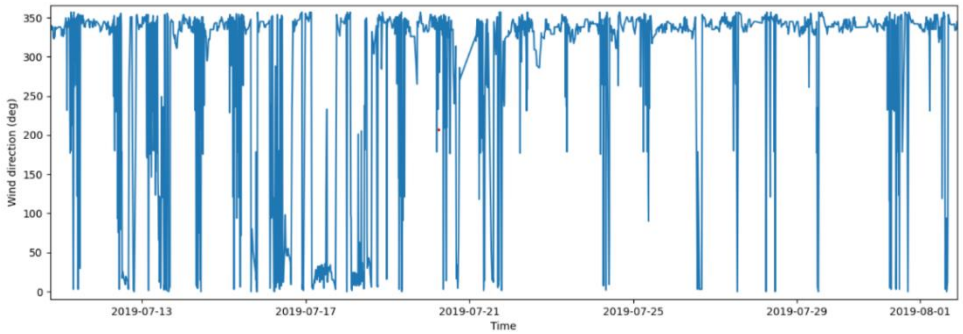
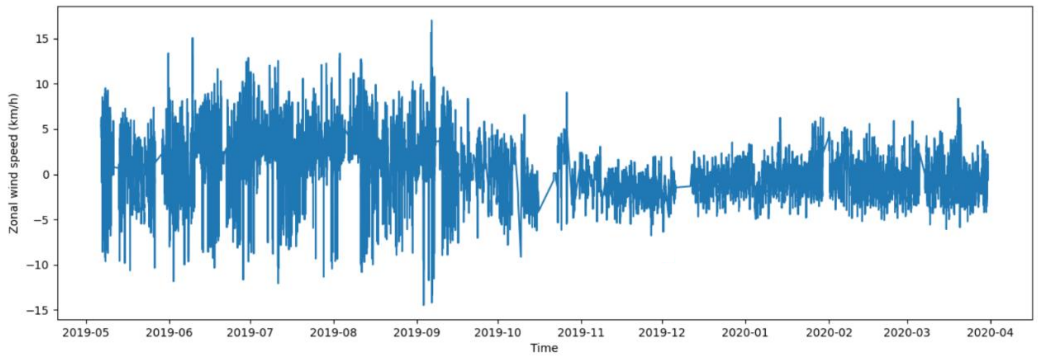
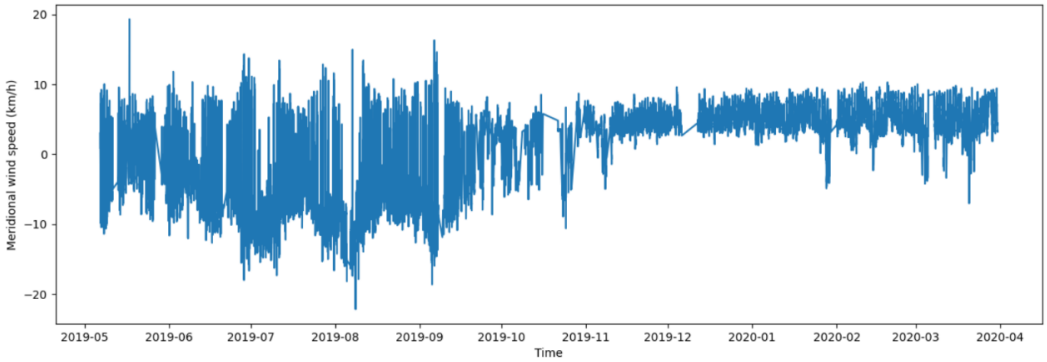


Figure 10. Input and output variables over time: a) active power, b) wind speed, c) wind direction, d) wind direction, detail.

a) Zonal wind speed



b) Meridional wind speed



c) Temperature

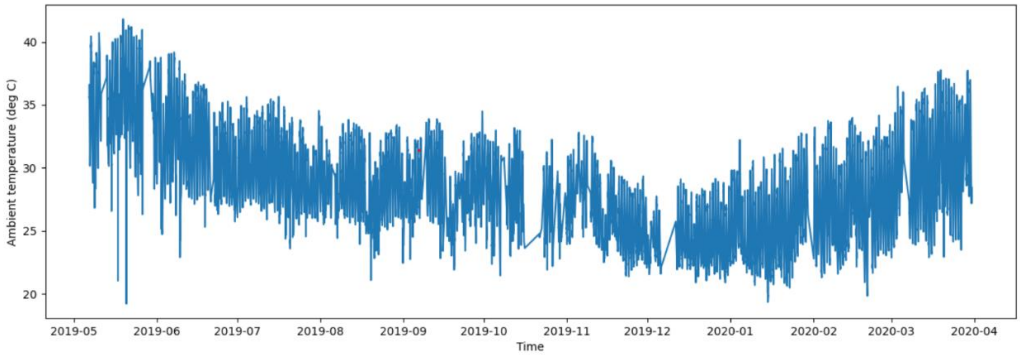


Figure 11. a) zonal wind componenent, b) meridional wind component, c) air temperature.

4.1.2. Variable relationships and feature selection

Relationships between output and input variables are shown in scatterplots (Figure 12), and correlations in Table 7a and 7b. Output power increases with wind speed (Figure 12a). Measurement points follow roughly S shape of a turbine power curve - only x and y axis switched. No further power can be produced after hitting max 1750 kWh/h. Wind speed is the most significant input feature for power production (corr = 0.930), while impact of wind direction (corr = 0.393) and temperature (corr = 0.127) is minor. Scatterplots are also in line with correlations. Meridional wind component is more significant than zonal wind speed component as expected since wind comes mainly from north (summer) or south (winter).

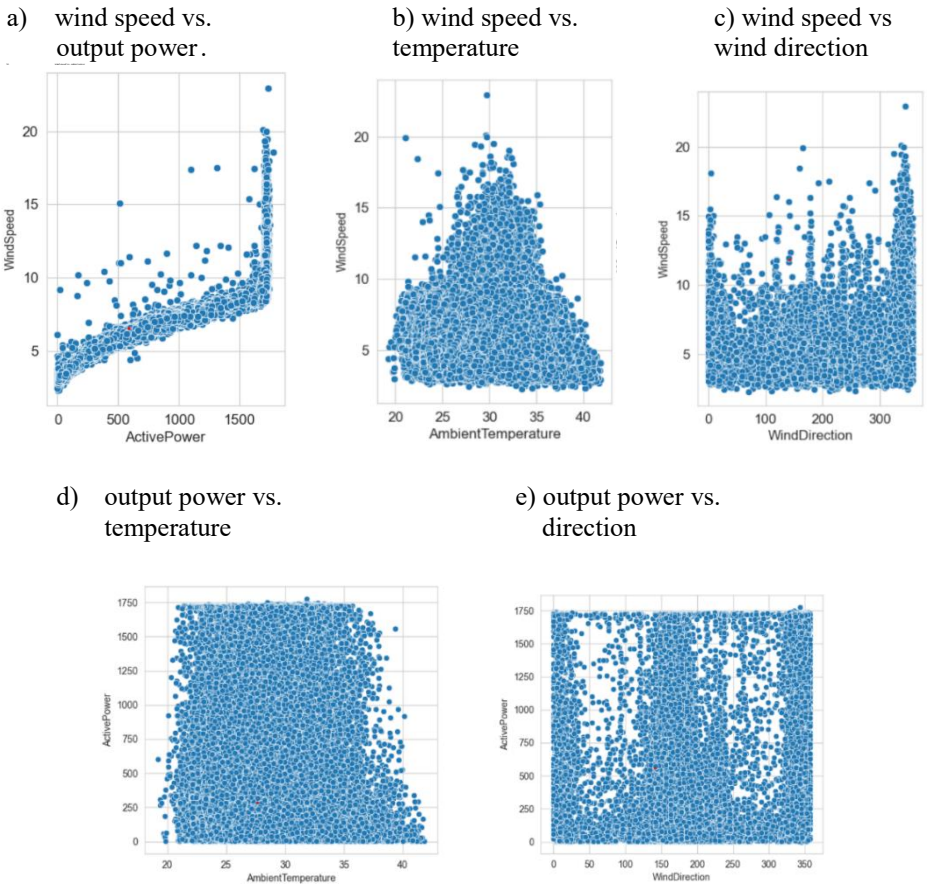


Figure 12. Variable relationships: a) wind speed vs. output power, b) wind speed vs. temperature, c) wind speed vs. wind direction, d) output power vs. temperature, e) output power vs. wind direction.

Table 7. Correlations between input variables and output power: a) wind speed and wind direction, b) zonal wind speed and meridional wind speed

Feature	Active power (kWh/h)	Air temperature (°C)	Wind direction (degrees)	Wind speed (km/h)
Active power (kWh/h)	1.0000	0.1271	0.3930	0.9304
Air temperature (deg C)	0.1271	1.0000	0.0930	0.1077
Wind direction (deg C)	0.3930	0.0930	1.0000	0.4067
Wind speed (km/h	0.9304	0.1077	0.4067	1.0000

Feature	Active power (kWh/h)	Air temperature (°C)	Zonal wind speed (km/h)	Meridional wind speed (km/h)
Active power (kWh/h)	1.0000	0.1271	0.2922	-0.4354
Air temperature (deg C)	0.1271	1.0000	0.0710	-0.3465
Zonal wind speed (km/h)	0.2922	0.0710	1.0000	-0.4531
Meridional wind speed (km/h)	-0.4354	-0.3465	-0.4531	1.0000

Wind speed and wind direction (alternatively meridional and zonal wind speed) are clearly important variables in wind power forecasting. Especially, zonal wind speed has much more impact compared to meridional wind speed. Temperature is in average around 30ºC. It varies between 20ºC – 40ºC and has minor significance. Temperature can be included in features, but since it is less important could be dropped as well.

4.2. EEM20 wind energy forecasting competition data

Conference on the European energy market 2020 hosted a competition on forecasting wind energy [85]. Data provided consists of three parts: gridded weather forecasts over Sweden, aggregate wind power production in the four different Swedish price regions, and a record of Swedish wind turbines and their locations. Weather and power production data covers years of 2000 and 2001, turbine data years 1885 – 2001.

4.2.1. Weather data

Weather data is given in NetCDF files – each file has hourly values for one day. NWP data is gridded over Sweden. In the competition, year 2000 data is historical, and year 2001 data is for weather forecasting.

Variables used are presented in Table 8. Most of them are used typically in wind power forecasting in general. However, wind gust speed has not seen to be used in any other wind power forecasts suggesting that it might be significant, especially in Scandinavia. Weather forecast variables are further given in 10 ensembles. Example of ensembles is a plot of temperature graph close to Luleå in January 2000 (Figure 13). Year 2000 and year 2001 (Appendix 1) seem to have fairly similar weather. These suggest that for longer term forecasting, it is beneficial to use historical data covering all seasons during years 2000 and 2001. Weather data of tree days is missing: 14th of May 2000, 26th of September 2000 ja 30th of July 2001.

Table 8. Weather forecast variables

Variable name	**Long name**	**Unit**
Temperature	Surface temperature (T2M)	K
Wind_U	Zonal 10 meter wind (U10M)	m/s
Wind_V	Meridional 10 meter wind (V10M)	m/s
WindGustSpeed	Wind gust	m/s
Pressure	Mean sea level pressure (MSLP)	Pa
RelativeHumidity	Screen level relative humidity (RH2M)	-
CloudCover	Total cloud cover (TCC)	-

Correlations of input variables have been calculated (Tables 9 and 10). Different weather parameters have typically only minor correlation with each other. Exception is zonal wind speed with wind gust speed (corr = 0.534 and 0.693). This might be explained by the fact that Scandinavian weather usually comes from west.

Table 9. Correlations of input variables in SE1 area year 2000. Pearson correlation

Variable	Tempera-ture	Zonal wind	Meridional wind	Wind gust	Air pressure	Relative humidity	Cloud cover
Temperature	1.0000	-0.0055	0.0737	0.2785	-0.0803	-0.5592	-0.1352
Zonal wind	-0.0055	1.0000	-0.2588	0.5341	-0.0947	-0.0943	-0.2887
Meridional	0.0737	-0.2588	1.0000	0.0487	-0.0933	0.2493	0.2528
Wind gust	0.2785	0.5341	0.0487	1.0000	-0.3078	-0.1551	-0.0563
Air pressure	-0.0803	-0.0947	-0.0933	-0.3078	1.0000	-0.2410	-0.3518
Relative	-0.5592	-0.0943	0.2493	-0.1551	-0.2410	1.0000	0.4816
Cloud cover	-0.1352	-0.2887	0.2528	-0.0563	-0.3518	0.4816	1.0000

Table 10. Correlations of input variables in SE1 area year 2001. Pearson correlation

Variable	Tempera-ture	Zonal wind	Meridional wind	Wind gust	Air pressure	Relative humidity	Cloud cover
Temperature	1.0000	-0.1469	0.1804	0.1066	0.1597	-0.4771	-0.0565
Zonal wind	-0.1469	1.0000	-0.0946	0.6928	-0.1924	-0.0144	-0.2166
Meridional	0.1804	-0.0946	1.0000	-0.0973	0.0264	0.1334	0.1446
Wind gust	0.1066	0.6928	-0.0973	1.0000	-0.2838	-0.1844	-0.0498
Air pressure	0.1597	-0.1924	0.0264	-0.2838	1.0000	-0.2420	-0.2257
Relative	-0.4771	-0.0144	0.1334	-0.1844	-0.2420	1.0000	0.4668
Cloud cover	-0.0565	-0.2166	0.1446	-0.0498	-0.2257	0.4668	1.0000

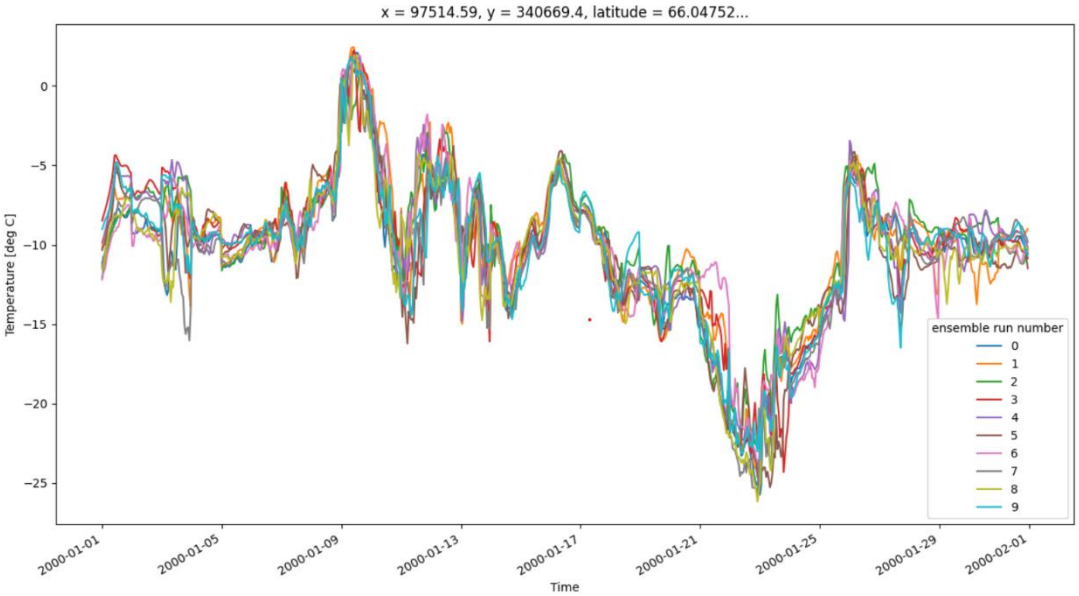


Figure 13. Temperature ensembles, January 2000. Location 66.07N 17.16E.

4.2.2. Wind power production data

Wind power production data is given as total values of four price regions in Sweden: SE1, SE2, SE3 and SE4 (Figure 15). Data is hourly covering years 2000 and 2001. As an example, first 3 rows from year 2000 are shown in Table 11. Wind power forecasting in the thesis is applied on RISE's ICE data center in Luleå in Sweden. Thus, focus is on the northernmost area SE1 called Luleå price region.

Table 11. Wind power production data, first 3 rows

Time	SE1 [MWh/h]	SE2 [MWh/h]	SE3 [MWh/h]	SE4 [MWh/h]
2000 0101 00:00:00	248.36751199999998	466.054477	764.836097	437.724342
2000 0101 01:00:00	238.10426500000003	449.537397	656.174814	427.607047
2000 0101 02:00:00	228.78177100000002	497.11267499999997	588.884484	386.638395

Total output power of wind turbines in SE1 area has increased over years (Figure 14). Two main reasons for this are: a maximum power of turbines has been increasing, and more wind farms have been built over the years. While the first turbines had max power clearly less than 1 MW (min only 20 kW!), in the end of 2001 it was already 4.2 MW. During years 2000 and 2001, 223 new wind turbines were installed.

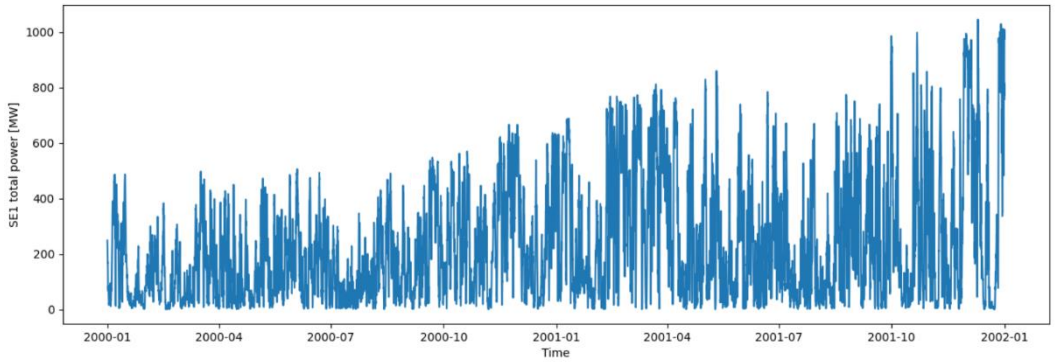


Figure 14. Output power of wind turbines in SE1 area, years 2000 - 2001

4.2.3. Wind turbine data

There are 472 wind turbines built in SE1 area during years 1980 - 2001. Data includes many turbine specific variables (Table 12). which are often used to model output power of a specific turbine, like for example in python's windpowerlib [36]. Locations of SE1 turbines are shown in Figure 15. In this thesis, installation dates and number of turbines are used in forecasting.

Table 12. Turbine variables

Variable name	Unit
Wind turbine ID	
Terrain height	m
Nacelle height	m
Rotor diameter	m
Max power	MW
Price region (SE1, SE2, SE4 or SE4)	
Installation date	
Longitude	180ºE - 180ºW
Latitude	90ºN - 90ºS

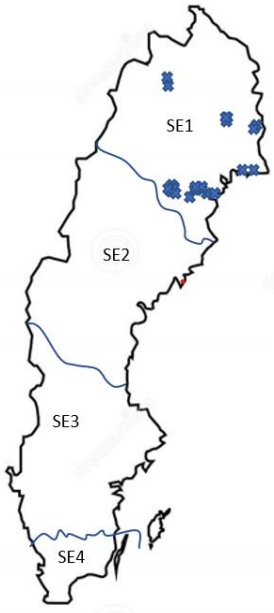


Figure 15. Electronics price regions and wind turbines SE1 area, Status end of year 2001. Source of borders [91].

4.2.4. Variable relationships ad feature selection

Produced output power is increasing over time (Figure 14), so it has been averaged to find a value per a wind turbine. Installation dates and number of wind turbines are main reasons for increasing trend and thus, have been taken into account in averaging. Result is more even production values over time (Figure 16).

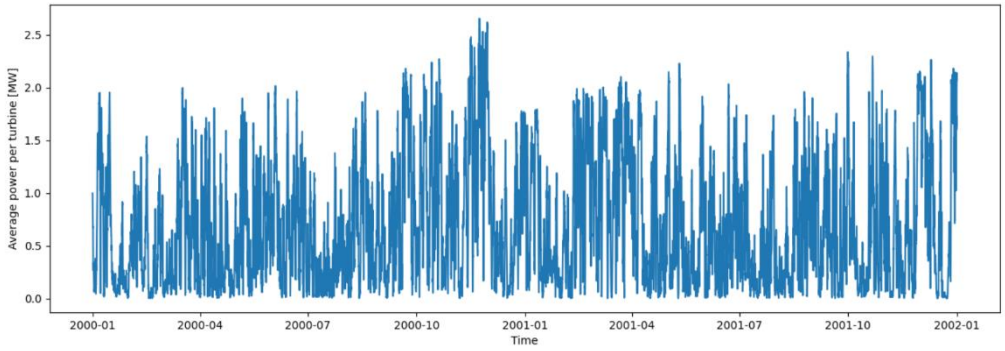


Figure 16. Averaged output power of wind turbines in SE1 area, years 2000 – 2001.

Correlations between input variables with output power were calculated for years 2000 and 2001 separately (Table 13). Both years give similar values. Results suggest that the most significant variables are wind gust speed and zonal wind speed. Zonal wind is more significant than meridional wind since Scandinavian weather typically comes from west. Air pressure seems to have some correlation.

Table 13. Correlations of input variables with produced averaged power in SE1 area. Pearson correlation

	Correlation	
Variable name	Averaged output power 2000	Averaged output power 2001
Temperature	0.0107	-0.1318
Zonal wind speed	0.5135	0.5107
Meridional wind speed	0.0830	-0.0896
Wind gust speed	0.7045	0.7237
Air pressure	-0.2012	-0.2653
Relative humidity	0.0971	0.0435
Cloud cover	-0.0750	-0.0176

Graphs in Appendix 2 show variable relationships with output power in January and July 2000. Relationships in January and July are clearly different. (See for example temperature and zonal wind speed.) Thus even more so than with Indian data, it is recommended that historical data for training is covering all seasons.

From variables in competition weather data temperature, zonal wind speed, meridional wind speed, air pressure and humidity are commonly used in wind power forecasts. Usually, wind speed and wind direction are given in data sets. They are used as such, or zonal and meridional wind components are calculated from speed and direction. Wind speed – especially zonal wind speed component in Nordic seems to have very significant correlation with produced power. Wind gust speed is not really used in any other forecasts. However, it seems to have very high correlation with produced power in the data suggesting that wind gusts have special importance in Scandinavian wind energy production. Air pressure has some, while cloud coverage

has very minor correlation. As a conclusion, wind gust speed, zonal wind speed and air pressure are “must” features to use in forecast, meridional wind speed and humidity are optional features. Cloud coverage can be included in feature set, but leaving it out should not have any real impact.

5. MODELLING AND FORECASTING EXPERIMENTS

Random forest regression model is the chosen model in the thesis. There are several reasons to support the choice. RFR is considered to have high accuracy in general. Both data sets are large. Especially Swedish data set has 28.1 GB of NWP data alone, and there is data about produced output power and turbines as well. Indian data set has a lot of missing measurement data.

5.1. India data

Data set is split into training, validation, and testing parts. May 2019, June 2019, July 2019, September 2019, October 2019, and November 2019 are used for training. August 2019 and December 2019 are used for validation. Model to be chosen is looked for by tuning hyperparameters. studying matching, avoiding overfitting, and finding suitable feature set.

5.1.1. Hyperparameters and matching

Hyperparameters need to be specified for model. Figure 17 shows averaged root mean square error as a function of number of trees for training data. As the number of trees increases then error decreases to a certain level. More trees means better accuracy, but at the cost of longer training time. Scikit learn's random forest regressor has default value of 100, which is fairly good compromise [92]. Value of 200 has been used in this thesis.

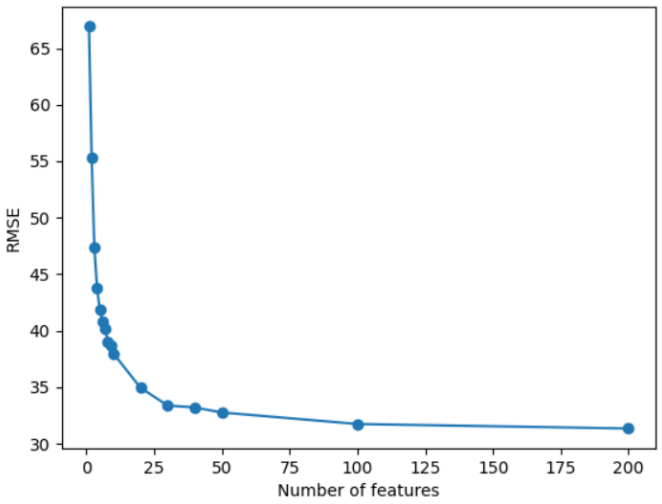


Figure 17. Root mean square error as a function of number of estimators (training data).

To find a criteria to stop algorithm, two hyper parameters were experimented: minimum number of samples at a leaf node, and maximum depth of a tree. Table 14 shows root mean square errors and mean average errors for some values. If no stopping criteria is used, model cleary overfits i.e. error with validation data is much higher than with training data. Using maximum depth seems to lead to higher errors, especially

when using zonal and meridional wind speed components as features. Thus, minimum number of samples at a leaf node is used in the thesis from now on. Chosen values of hyperparameters are shown with green font in the table: 200 estimators and minimum samples per leaf 10.

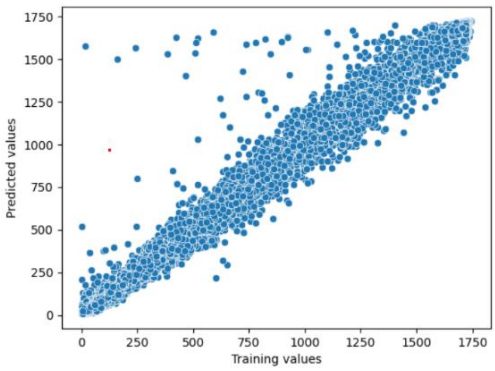
Table 14. Root mean square error and mean average errors with stopping criteria. a) wind speed and wind direction as features, b) zonal wind and meridional wind as features

a)	**Training data**		**Validation data**	
Hyperparameters	**RMSE**	**MAE**	**RMSE**	**MAE**
n_estimators = 200	31.8600	17.9566	80.2347	50.8749
n_estimators = 200, min_samples_leaf = 10	71.5714	39.4200	73.1052	53.3929
n_estimators = 200, max_depth = 6	77.7645	46.0537	72.3981	47.8975
n_estimators = 200, min_samples_leaf = 10, max_depth = 6	79.6605	46.1363	72.1846	51.1235

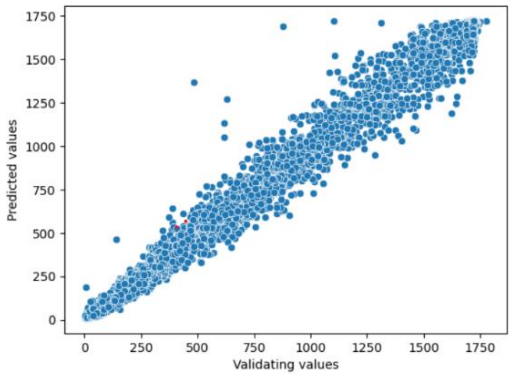
b)	**Training data**		**Validation data**	
Hyperparameters	**RMSE**	**MAE**	**RMSE**	**MAE**
n_estimators = 200	32.3517	18.3010	81.6839	61.6139
n_estimators = 200, min_samples_leaf = 10	74.6841	41.7347	80.9574	51.2819
n_estimators = 200, max_depth = 6	140.8454	98.7693	132.7308	110.4251
n_estimators = 200, min_samples_leaf = 10, max_depth = 6	141.7282	98.7070	132.6970	110.7214

Graphs in Figure 18 visualize how well forecasted power values match with actual power values with hyperparameter values chosen above (200 estimators and minimum samples per leaf 10). Some power values are forecasted too high. This happens more with training than validating data. Other hyperparameter values were also looked at, but the matching with the chosen ones was the best one.

a) training, wind speed & direction



b) validating, wind speed & direction



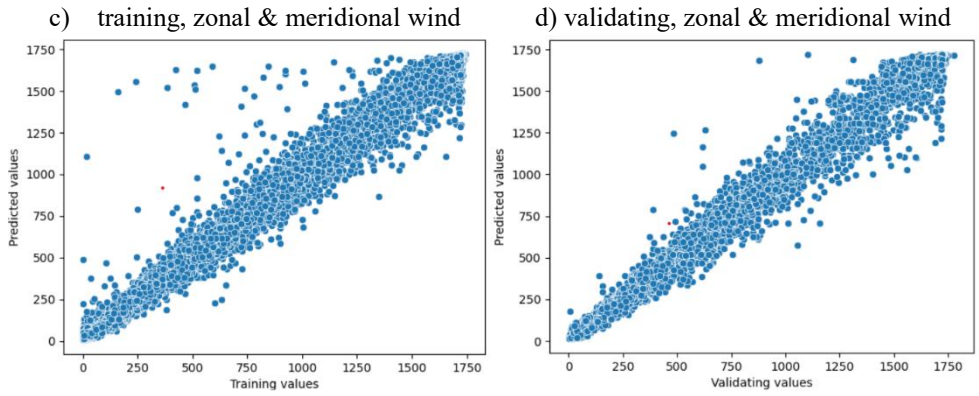


Figure 18. Matching of forecasted and actual power values. With wind speed, wind direction and temperature features: a) training data, b) validating data. With zonal and meridional wind speed and temperature features: c) training data, d) validating data. (n_estimators=200, min_samples_leaf=10).

5.1.2. Impact of features

Feature importances were calculated for variables (Table 15) [92]. Results are in line with correlation values (Table 7) and scatterplots (Figure 19) reported earlier showing that wind speed (fi=0.998) and meridional wind speed (fi=0.86) are defining most of produced power.

Table 15. Feature importances. a) wind speed and wind direction, b) zonal and meridional wind speed (n_estimators=200, min_samples_leaf=10)

a)

Feature	**Feature importance with wind speed and wind direction**
Air temperature (deg C)	0.00171
Wind speed (km/h)	0.99810
Wind direction (deg)	0.00018

b)

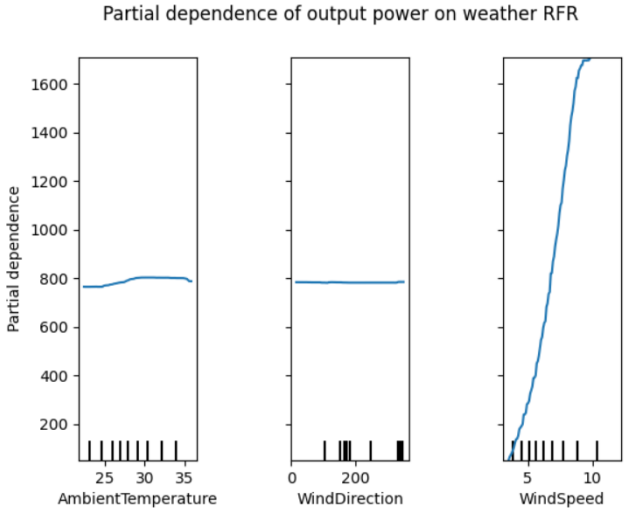
Feature	**Feature importance with zonal and meridional wind speed**
Air temperature (deg C)	0.00197
Zonal wind speed (km/h)	0.13530
Meridional wind speed (km/h)	0.86273

Partial dependence plot shows the marginal effect one or two features have on the predicted outcome of a machine learning model [93, 94]. It can show whether the relationship between the target and a feature is linear, monotonic, or more complex.

Plotted partial dependencies (Figure 19) confirm conclusions already made with correlations and feature importances. Wind speed has the highest impact on produced output power. Relationship follows gently S curve similar to a power curve of a turbine. Meridional and zonal wind speeds have U shape dependance. Meridional component has clearly more impact than zonal one. Since, in the location wind comes

either form north or south, there are practically no samples around zero value. Temperature has only trivial role in forecasting in both cases.

a)



b)

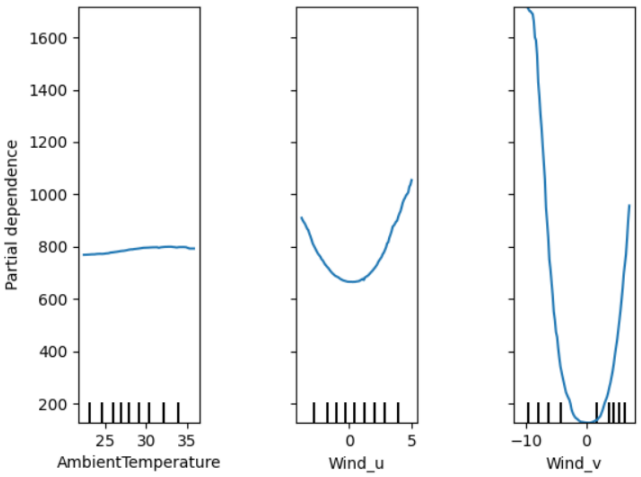


Figure 19. Partial dependencies of features: a) wind speed and wind direction used, b) zonal wind and meridional wind used.

Reducing number of features may decrease computing load. Thus, impact of smaller feature set on root mean square error has been calculated (Table 16). Less features means increased error. Zonal wind speed has s high feature importance of 0.135 (Table 15), so dropping it makes RMSE very high (around 240). Full feature set has been selected in both cases. Selections are marked with green font in Table 16.

Table 16. Root mean square errors with different feature sets

Features	**RMSE training data**	**RMSE validating data**
WindSpeed, WindDirection, Temperature	71.5658	73.1734
WindSpeed, WindDirection	74.2488	75.5440
WindSpeed	79.7375	79.6360

Features	**RMSE training data**	**RMSE validation data**
Wind_V, Wind_U, Temperature	74.6706	80.6191
Wind_V, Wind_U	76.9627	82.3019
Wind	243.5211	239.0756

Figure 20 shows matching of forecasted and actual power with different feature sets. It is seen, that reduced feature set deteriorates matching. In first case (wind direction, wind speed and temperature) using wind speed as only feature forecasts produced power fairly well. In the second case (meridional wind speed, zonal wind speed and temperature), meridional wind as only feature leads to poor matching. Based on matching it is chosen to use full feature set in both cases here as well.

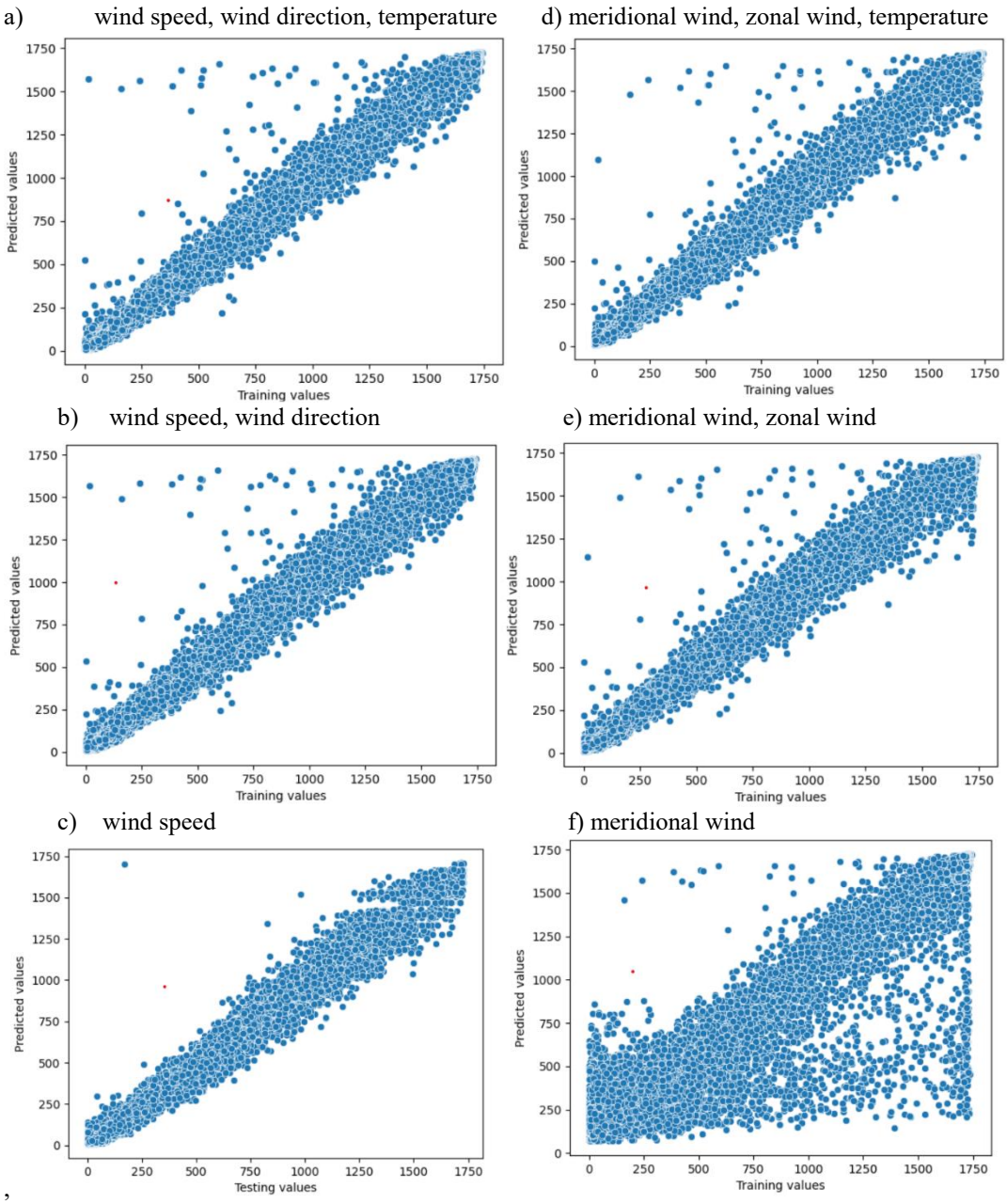


Figure 20. Example about matching of forecasted and actual power values with different feature sets. a) Wind speed, wind direction and temperature, b) wind speed and wind direction, c) wind speed, d) meridional wind, zonal wind and temperature, e) meridional wind and zonal wind, f) meridional wind. All training data. (n_estimators=200, min_samples_leaf=10).

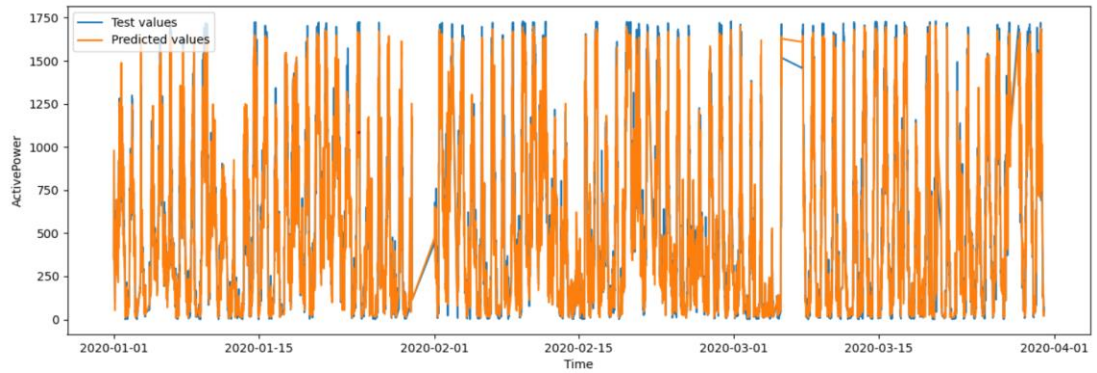
5.1.3. Forecasting experiments

In forecasting experiments full feature sets have been used - wind speed, wind direction and temperature in first case, and zonal wind speed, meridional wind speed and temperature in second case. Hyperparameters for RFR model are 200 estimators and minimum samples of leaf 10. Figures 21 and 22 show forecasting results for testing values. Forecasted values follow actual values pretty well, as can be seen in the detailed graphs.

Comparison of matching to training and validation data (Table 17) shows that root mean square error is somewhat higher with both cases. There are only some testing values, which are forecasted to higher output power than expected (Figure 23).

In addition to RFR, also decision tree regression (DTR) was looked at. DTR overfits very easily. Different hyperparameter values were able to only alleviate the problem slightly. The worst case was when default value of 1 for hyperparameter *min_samples_leaf* was used. It caused model to overfit completely.

a)



b)

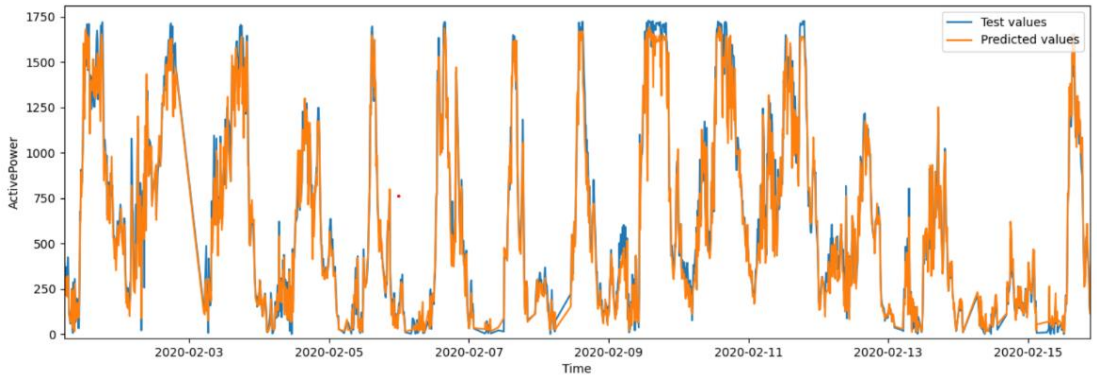
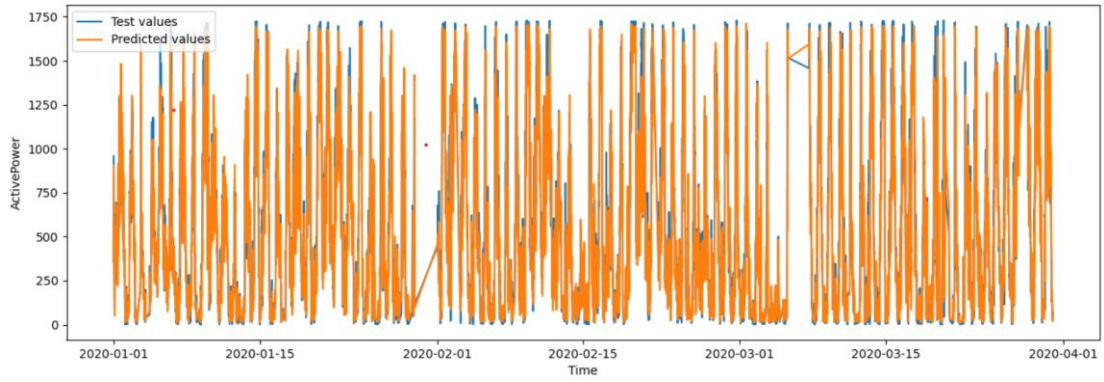


Figure 21. a) Forecasting with wind speed, wind direction and temperature and b) detail of the graph.

a)



b)

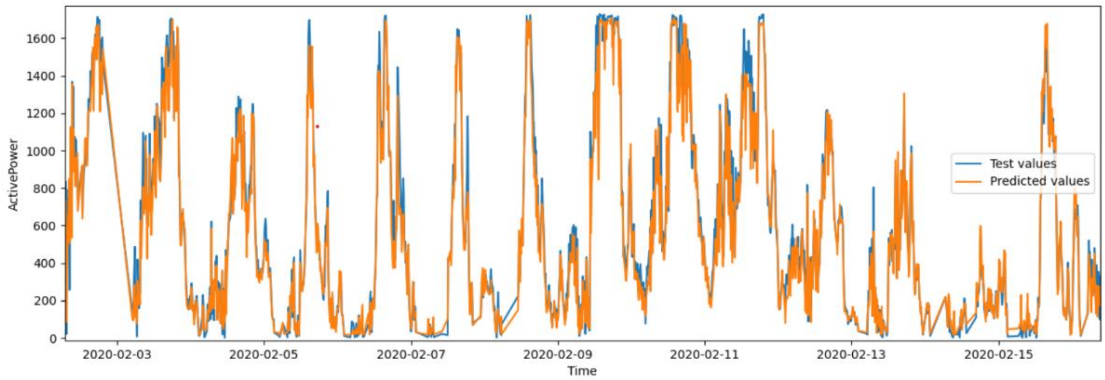


Figure 22. a) Forecasting with zonal wind speed, meridional wind speed and temperature and b) detail of the graphs.

Table 17. Root mean square error comparison

Features	**RMSE Training data**	**RMSE Validation data**	**RMSE Testing data**
Wind speed, wind direction, temperature	71.5658	73.1734	75.3089
Meridional wind, zonal wind, temperature	74.6706	80.6191	84.1761

a) b)

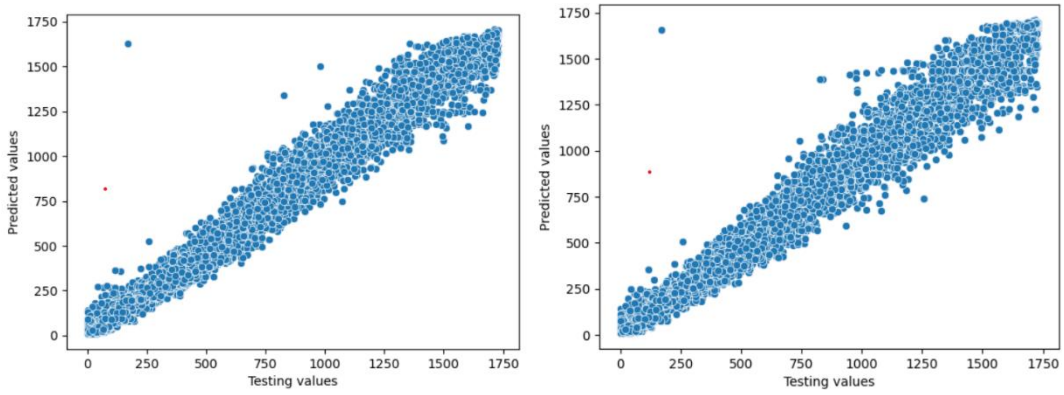


Figure 23. Matching of testing and forecasted power: a) with wind speed, wind direction and temperature, b) with meridional wind speed, zonal wind speed and temperature.

5.2. EEM20 competition data

5.2.1. Nature of data

Competition data is originally meant to forecast total wind energy production in Sweden. However, in the thesis, data is applied to estimate production of one wind turbine. Using the data differently causes imprecisions. Produced power is given as total value for price regions SE1, SE2, SE3 and SE4. Power of region SE1 is divided by number of wind turbines at certain times to get an average number for a wind turbine. Each wind turbine has been built at different times and locations, has different features like maximum power, etc. (Table 12). Weather data is given on defined grid points. Geographical location close to Luleå has been chosen for weather variables (66.07N, 17.16E). Some other locations were tried as well but they resulted in higher errors.

Although competition data is aggregate in nature, there are clear benefits to use it. Wind energy production depends a lot on local weather conditions. Data allows to estimate relative importances of different weather parameters, specifically in northern Sweden. For example, Indian data could not be applied to Scandinavia due to very different climate.

Weather data from year 2000 was used as historical data in competition. It covers whole year with all seasons, and thus, is a good choice to be used to train the RFR model. Year 2000 data has been used to train forecasting model in the thesis. January, February, March, May, June, July, September, October, and November 2000 were used for training. April, August, and December used for validation. Year 2001 data is used for testing.

5.2.2. Hyperparameters and matching

Number of estimators used here is 200, the same as with Indian data. Value is a compromise between accuracy and computing time as already explained in Section 5.1.1.

To study impact of stopping criteria, two hyperparameters were looked at: minimum number of samples at a leaf node and maximum depth of a tree. Root mean square errors and matching were computed (Table 18, Figure 24, and Figure 25). As errors and matching plots show optimal hyperparameter values are hard to find. Without stopping criteria model overfits (first row of Table 18 and Figure 24a). With maximum depth, the model seems not to be able to forecast low/high values correctly (Figure 24c). Model on 2nd row with min_samples_leaf seems to be best choice from overfitting point of view. model on 2nd row has been chosen (green font). The hyperparameters selected are the same as with India data: n_estimators = 200, min_samples_leaf=10.

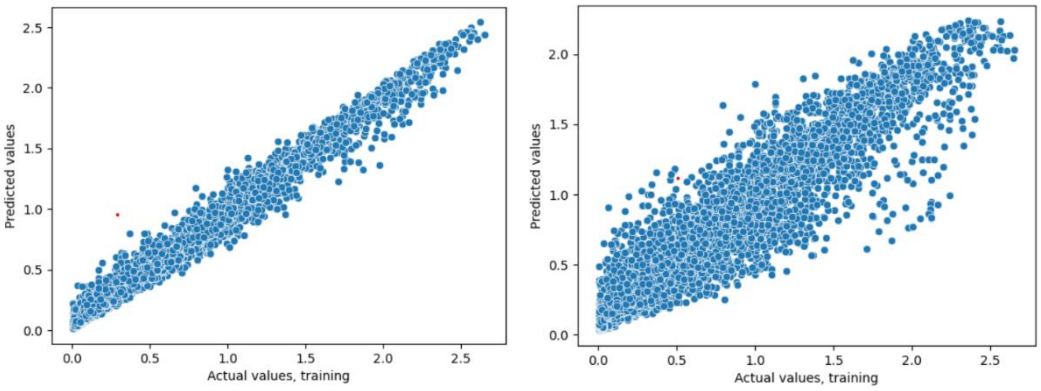
Major reasons for the matching problems especially with validation data are most probably due to the nature of data as explained in Section 5.2.1.

Table 18. Root mean square error and mean absolute errors with stopping criteria. Default parameter values for stopping are: min_samples_leaf=10, max_depth=None

	Training data		Validation data	
Hyperparameters	**RMSE**	**MAE**	**RMSE**	**MAE**
n_estimators = 200	0.0808	0.0540	0.3435	0.2746
n_estimators = 200 min_samples_leaf = 10	0.2110	0.1489	0.3469	0.2762
n_estimators = 200 max_depth = 6	0.3060	0.2304	0.3485	0.2769
n_estimators = 200 min_samples_leaf = 10 max_depth = 6	0.3104	0.2329	0.3472	0.2764

a) training, 200 estimators

b) training, 200 estimators, min samples leaf 10



c) training, estimators 200 max dept 6

d) training, estimators 200, min samples leaf 10, max depth 6

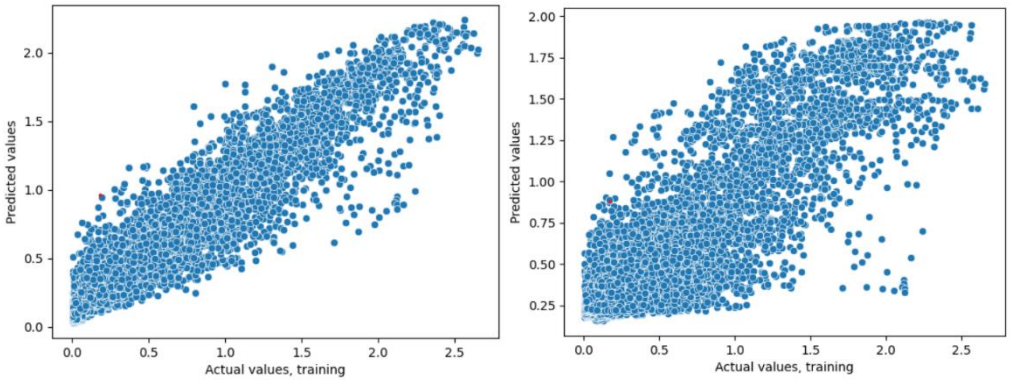


Figure 24. Matching of forecasted and actual power values with training data: a) training, n_estimators=200, b) training, n_estimators=200, min_samples_leaf=10, c) training, n_estimators=200, max_depth=6, d) training, n_estimators=200, min_samples_leaf= 10, max_depth=6.

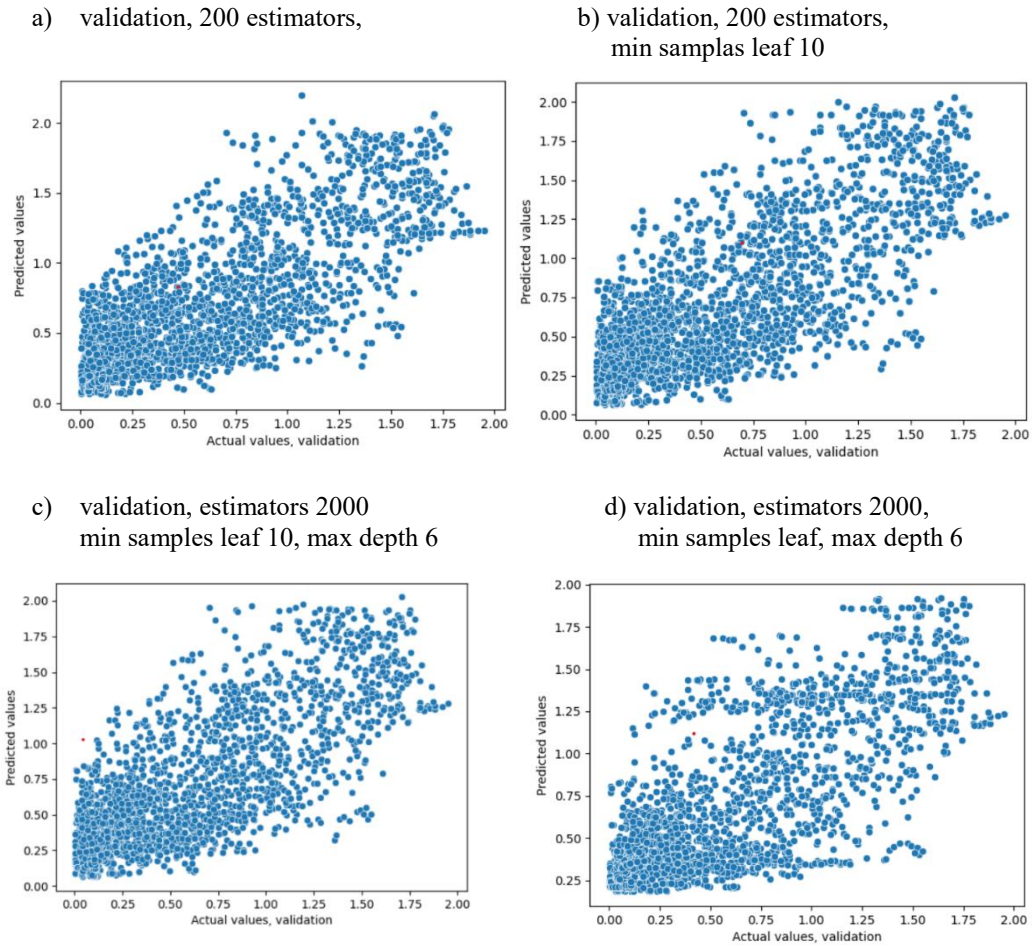


Figure 25. Matching of forecasted and actual power values with validation data: a) validation n_esitmators=200, b) validation, n_estimators=200, min_samples_leaf=10, c) validation, n_estimators=200, max_depth=6, d) validation, n_estimators=200, min_samples_leaf=10, max_depth=6,

5.2.3. Impact of features

Feature importances are shown in Table 19. Wind gust speed clearly stands up having importance of 0.55. In line with this, its correlation is also high 0.7 (Table 13). Wind gust is a special feature, since it has not been used in any other wind power forecasts than in the EEM20 competition. Other features having fair importance are zonal wind speed and temperature. As already explained, Scandinavian weather is typically coming from west. This is clearly different from Indian data, where northern and southern winds are dominant confirming presumption that local weather data has to be used in each forecasting case. Relative humidity, air pressure and meridional wind speed have roughly the same importance. Cloud cover is insignificant.

Calculated partial dependencies are shown in Figure 26. Graphs show that wind gust speed and zonal wind speed have high dependence. This is in line with feature importances, and correlations already computed. Wind gust speed has especially high values. Most features have steadier dependences. Exception is cloud cover, which has very minor dependance. This suggests that cloud cover is potential feature to drop from feature set. Note that both feature importances and partial dependencies depend on the

used model. Thus, if any of the hyper parameters is changed, calculations need to be rerun.

Table 19. Feature importances. (n_estimators=200, min_samples_leaf=10)

Feature	Feature importance
Wind gust speed	0.54766
Temperature	0.12702
Zonal wind speed	0.10893
Relative humidity	0.07742
Air pressure	0.06357
Meridional wind speed	0.05417
Cloud cover	0.02124

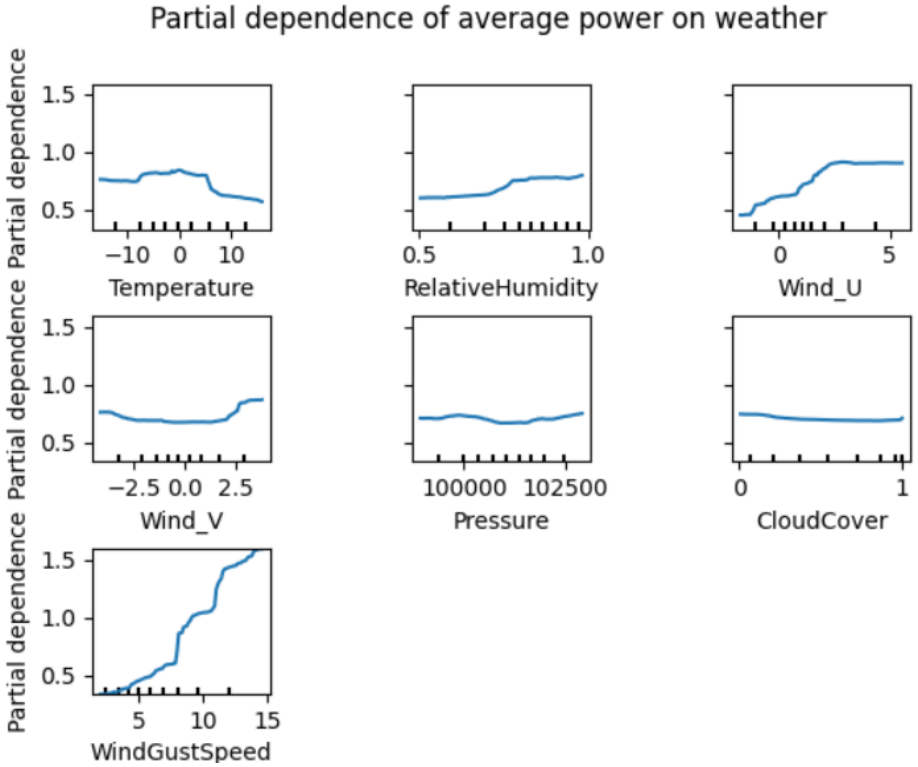


Figure 26. Partial dependencies of features.(n_estimators=200, min_samples_leaf=10)

Advantage of using a smaller feature set is a lower computing load, so possibility to reduce the set is looked at. Root mean squares for different feature sets are shown in Table 20. Dropping cloud cover impacts only slightly on error and matching (blue font). In line with this, cloud cover also has the lowest importance and correlation. If features are dropped further in the order of importance, error gets higher. Two examples of matching graphs are shown in Figure 27 for 7 features and 6 features. Matching in the figure is in line with RMSE errors, confirming minor importance of cloud cover. Also smaller feature sets were looked at, but errors got higher and matching worse as expected.

Table 20. Root mean square errors with different feature sets. (n_estimators=200, min_samples_leaf=10)

Features	RMSE training data	RMSE validating data
wind gust, zonal wind, meridional wind, temperature, humidity, pressure, cloud cover (all 7 features)	0.2110	0.3469
wind gust, zonal wind, meridional wind, temperature, humidity, pressure (6 highest importances, cloud cover dropped)	0.2121	0.3487
wind gust, zonal wind, temperature, humidity, pressure (5 highest importances)	0.2186	0.3552
wind gust, zonal wind, temperature, humidity (4 highest importances)	0.2464	0.3712
wind gust, zonal wind, temperature (3 highest importances)	0.2805	0.3947

Based on errors and matching all 7 features give the most accurate result. Cloud cover feature can be dropped from the set without any real impact. Thus, there are two useful feature sets: a) all 7 features (green font) or b) 6 features, cloud cover dropped (blue font). However, all 7 features have been chosen in the experiments, since set gives the smallest error (green font).

a) 7 features b) 6 features, cloud cover dropped

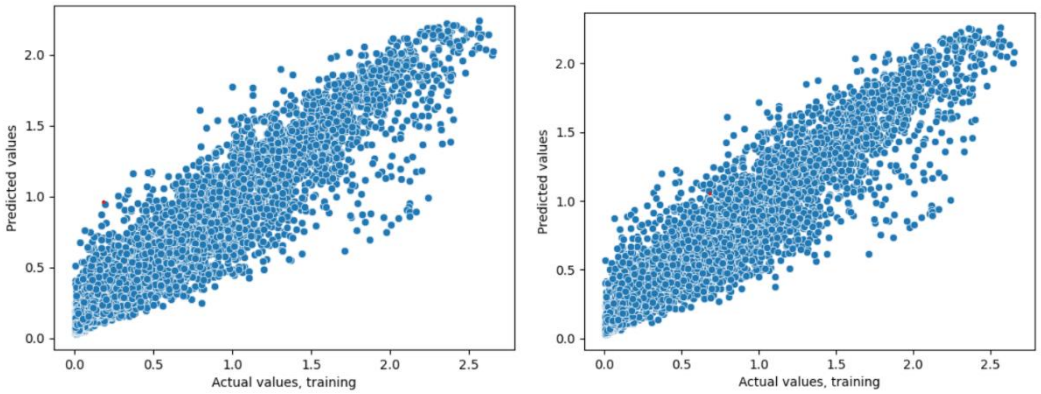


Figure 27. Examples about matching of forecasted and actual power values with different feature sets. a) 7 features, b) 6 features, cloud cover dropped. Training data used. (n_estimators=200, min_samples_leaf=10).

5.2.4. Forecasting experiments

Forecasting with the model is done so, that forecasted weather parameters are fed to the input of the trained model. Forecasted production power values are then calculated with the model. Hyperparameters for RFR model are 200 estimators and minimum samples of leaf 10. All seven features are used.

Forecasting experiment is shown in Figure 28. Forecast is fairly accurate at some time points, but not able to predict all the time. Matching plot (Figure 29) shows also more spread distribution compared to training and validation plots (Figures 24 and 25). The same can be seen in RMSE values (Table 21). Error with testing values is much

higher than error with training or validation values. Sources of errors are most probably due to the nature of data. See Section 5.2.1.

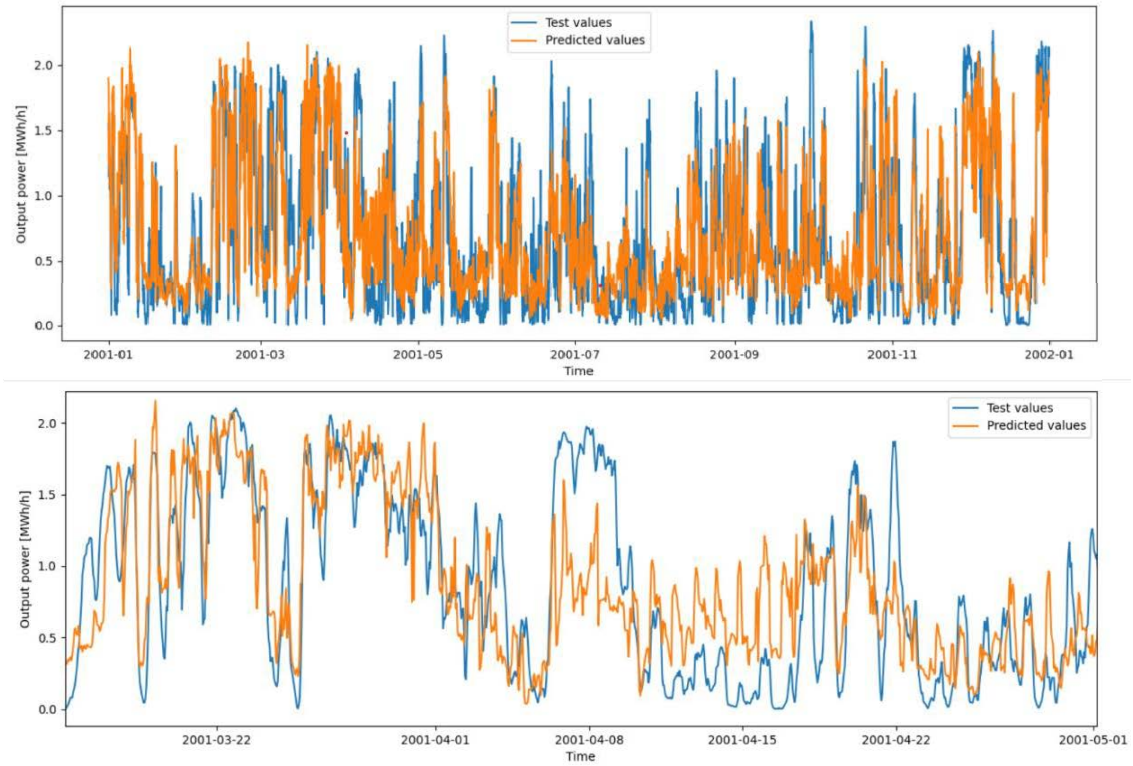


Figure 28. a) Forecasting with all 7 features and b) detail of the graph.

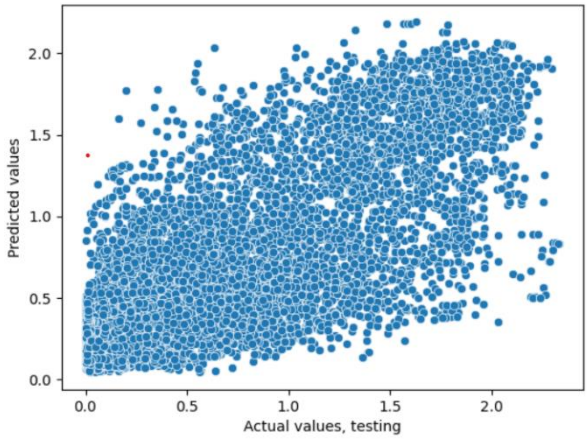


Figure 29. Matching of testing and forecasted power with all 7 features.

Table 21. Root mean square error comparison

Features	**RMSE Training data**	**RMSE Validation data**	**RMSE Testing data**
all 7 features	0.2110	0.3469	0.4016

6. APPLICATION TO A DATA CENTER

6.1. RISE data center

Research Institutes of Sweden (RISE) has an ICE data center in Luleå Sweden for research purposes [95, 96]. It is used to study facilities including power and cooling, and IT; and how these co-operate to achieve the best possible system performance [97]. Figure 30 shows schematic drawing of facility components.

Microgrid controller (FIMER MGS100) manages power sources and delivery of power to IT and cooling. Controller also takes care of charging the batteries. There are currently 3 power sources: national grid i.e. so called brown energy, batteries (NorthStar NSB 100FT BLUE+, 30 kWh), and solar panels (Jinko Solar Eagle MX, 10 kW). There is reservation for a 10 kW wind turbine.

Container module includes IT equipment and computer air room handler (CRAH). IT equipment has two 42U racks both containing 38 servers (Dell PowerEdge R430) and two switches (Dell N1548). Idle and maximum load of servers are 6 kW and 12 kW. Since extra energy in a form of heat is produced by the servers, container module has to be cooled down.

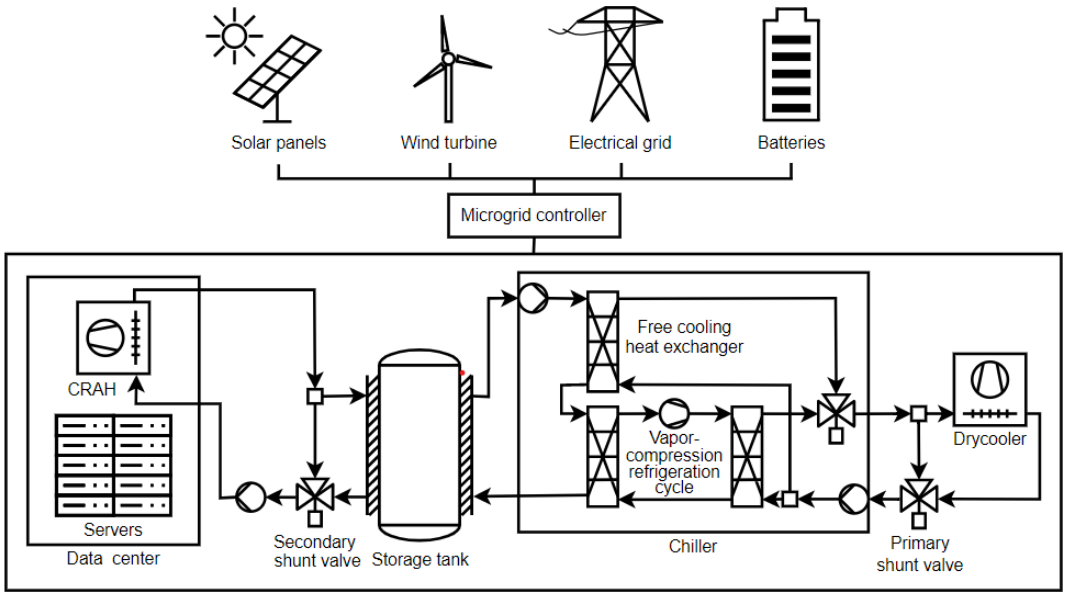


Figure 30. A schematic drawing of facility components in RISE ICE data center. Figure drawn by RISE.

CRAH has a fan, which is forcing air from hot to cold aisle through heat exchanger in the module. Thermal energy storage tank has 2000 liter of water. Cool water flows from tank to CRAH where water absorbs heat from exchanger and then flows back to the tank. Water in the tank is further cooled down in chiller (BlueBox Tetris W Rev FC/NG). Chiller can function in chiller mode, free cooling mode, or mixed mode. Chiller mode uses vapor – compression refrigeration. Water – glycol mixture is used as fluid. Glycol mix moves to heat exchanger in dry cooling tower and is cooled by colder outside air. In case outside temperature is low enough, free cooling mode is

adequate, and no chiller mode is needed. Cooling system consumes 9 kW of power. It is able to transfer heat power up to 32 kW, but in practice only about 10 kw is needed.

Controller chooses one of the sources at a time. Aim has been for data center to be able to operate for 3 hours on its own in the case of power break in a national grid [98]. Batteries are the main back-up system, while solar panels give additional support. Power and cooling system has been modelled with Matlab and Simulink.

6.2. Test input variables

Forecasting model has been trained with EEM20 competition data from year 2000. Test input variables are basically the same as in EEM20 competition. There are differences in measurement units though, so input variables of test data need to be adjusted to match the ones in training data (Table 22). Note that zonal wind speed and meridional wind speed had to be calculated from wind speed and wind direction. Test variables come from Swedish Meteorological and Hydrological Institute (SMHI) open database. Data covers 10 days from 2nd of March to 12th of March 2021 in Luleå [98].

Table 22. Input variables for testing with trained model and required adjustments

Feature	**input variables**	**for trained model**
temperature	ºC	ºK
relative humidity	% , 0 – 100	0 - 1
wind speed	m/s	zonal wind speed m/s and meridional wind speed m/s
wind direction	degrees from north	
wind gust speed	m/s	m/s
air pressure	hPa	Pa
cloud cover	0 – 8	0 - 1

Figure 31 shows wind speed of test input variables. Wind gust speed follows the same shape as wind speed. Wind speed and wind gust speed are the most important features in predicting output power as explained in Section 5.2.3. Prediction frequency of input variables is 1 hour in the beginning of time-series increasing to 12 hours in the end. Graphs of all input variables for testing as well as calculated zonal wind speed and meridional wind speeds are shown in Appendix 3.

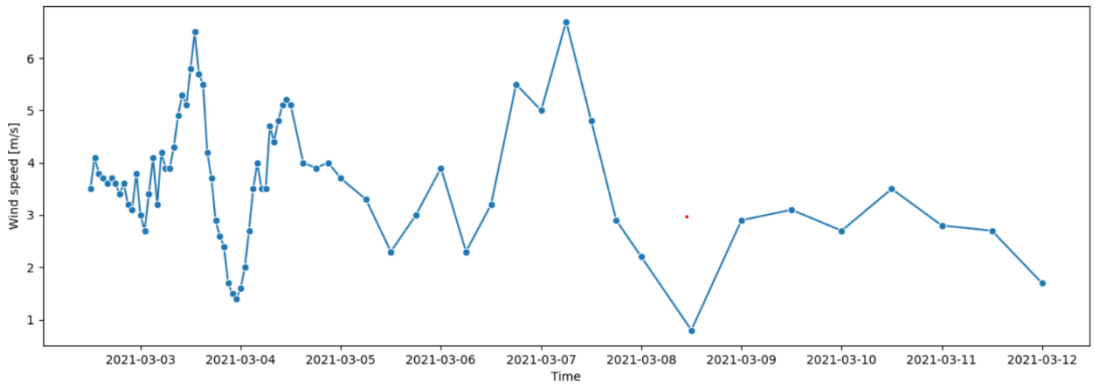


Figure 31. Wind speed of testing data.

6.3. Training and testing code

Forecasting code for RISE has been split into two parts. Flow charts of both parts are shown in Figure 32. First part is the training of the RFR model with competition data. All 7 features have been used in training. Output power is normalized by dividing a mean power by a maximum value before actual training. Trained model is saved to a file.

Second part does the actual testing and forecasting. This part utilizes forecasted weather variables over the time period where output power forecasting is looked for. The test data covers 10 days. Trained model and test input variables are read in. After adjusting variables to training data, forecasting is performed with RFR model. Forecasted values are saved to a CVS file. RISE simulation model runs the second part of the code and then uses the result stored in CVS file in its own simulation.

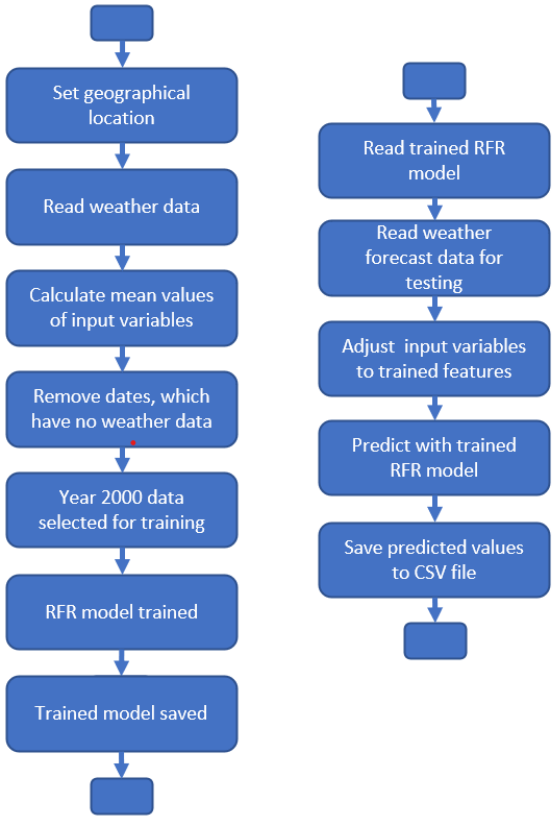


Figure 32. Flow charts of training code (left) and testing code (right).

6.4. Forecasting experiment

Figure 33 shows the forecasted output power. Output power follows fairly well a general shape of the wind speed (Figure 31) as expected. Towards the end following is less accurate. On the other hand, prediction frequency is getting sparser, and further in time weather forecast is getting less accurate.

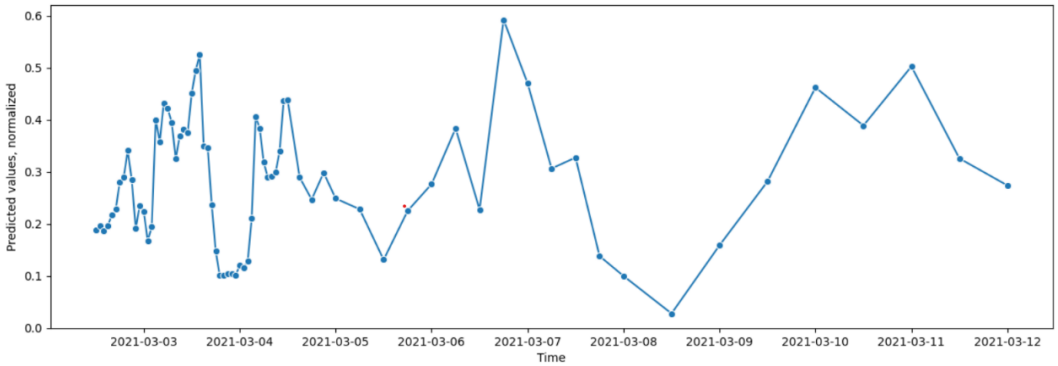


Figure 33. Normalized predicted output power for test weather data.

It is worth noticing that although wind speed is the most important feature in prediction it is not the only factor. There are 5 other features impacting prediction as well. Additionally, training data has sources of inaccuracies as explained in Section 5.2.1.

6.5. Application and sustainability

Sustainability is becoming one of the keys issues in data center business [28]. Using renewable energy sources is a way to produce sustainable energy. More wind energy is produced in winter than in summer (see Section 2.1.1). On the other hand, in Nordic countries there is clearly more solar energy available in summer than in winter. Using combination of wind and solar energy could thus reduce impact of seasonal variety. Wind energy can also increase data center's resiliency against costly electricity outages.

RISE is planning to use forecasting result to find optimal relative sizes of renewable energy sources for ICE data center. This is done by using the RFR model developed in the thesis combined with RISE's Matlab and Simulink simulations.

7. DISCUSSION

Applicable data sets for wind power forecasting were challenging to find. Access to private data sets would in practice require partnership with the other party, a company or research institute, having wind turbines. Commercial data is most probably too sensitive for open access. While looking for literature about wind power forecasting, no articles were noticed about the subject concerning Sweden. Perhaps, research in this area in Sweden is not public.

Two open data sets were found: one from India and another from Sweden. Data from two different countries gave perspective about importance of climate in wind energy forecasting. Due to different weather conditions also feature sets are different. In addition, same features can have different relative importances. For example, a feature which has high importance in Sweden might be irrelevant in India. Wind gust is such a feature. All this means that a model trained with Indian data is not applicable in any Swedish location and vice versa.

Indian data is from one wind turbine. This allows to develop statistical model accurately in comparison to Swedish data where total output power of 472 wind turbines had to be averaged. Time range covers 2.5 year period, but because there was a lot of data missing, only one year was usable. Especially in the beginning of time-series, most measurement data was missing. India has two seasons, where wind is blowing from north half a year, and from south the other half. Longer time range would have been useful for the training data to cover seasonality properly.

Usable feature set in Indian data was quite small. Most variables in the data set were about the turbine itself. Wind speed, wind direction and temperature were used. Temperature did not have any significant importance, so it could have been dropped out from feature list. Temperature ranged from 20 to 40 degrees; average on 30 degrees. Two or maximum three features is small feature set compared to sets used in literature in general. However, the local climate was so straightforward, that small feature set seemed to be enough to forecast output power quite accurately.

Swedish data set comes from EEM20 forecasting competition. Goal of the competition was to forecast total wind power production in Sweden, while target of the thesis is to estimate forecast for one wind turbine near Luleå. Thus, competition data had to be transformed and modified for the thesis. Weather variables are given in grid points over Sweden. A point as close as possible to Luleå was chosen. Error rising from a location difference is estimated to be small. Additionally, NWP data had 10 ensembles, which gave more accuracy.

Total output power for four price regions in Sweden were given in the competition. Production values for SE1 area were naturally used. Here are the major sources of error compared to a data set of only one wind turbine or one wind farm., Totally 488 wind turbines have been built in SE1 area during years 1885 – 2001. Out of those, 233 were installed during years of competition data 2000 and 2001. Turbines were built on different locations across SE1 area, they had different maximum output power values and different installation dates. Output power of one wind turbine was estimated by calculating average of SE1 at certain time points. Number of turbines and their installation dates were included in averaging.

Competition data allowed to study wind power forecasting from interesting aspects, although it had to be modified for the thesis. Weather data included SE1 area where Luleå is located, so climate was the right one. It was suitable for estimating the relative importances of weather variables in Northern Sweden. For example, it showed how

important feature wind gust is in the area. Stormy weather might cause so much wind power production, that price of electricity can be negative at times. One year of data was used to train the model. Covering all seasons improves forecasting accuracy.

The data available both in India and Sweden cases support statistical modeling. Physical models would require more information, for instance about surroundings and terrain. Random forest regression is a statistical model, which handles well large data sets. It is considered to be highly accurate. Overfitting can be dealt with by tuning hyperparameters carefully. RFR performed well in forecasting experiments. India case was more straightforward since data covered only one wind turbine. Forecasting with Swedish competition data did also well, considering especially approximations made with the data.

Forecasting model has been developed for RISE's ICE data center. Currently RISE has solar panels, but no wind turbines in their microgrid. Developed model will be used by RISE to study optimal relative sizes of wind and solar energy production. Using two renewable energy sources increases data center's resiliency and balances impact of seasonal variations. Final testing of the model will be done in RISE's simulation environment.

Major limitation in the research are related to both data sets. As already discussed, longer time-series in Indian data would cover seasonal variabilities better for training. Situation of Swedish data set is more complex, since it has to be modified. Ideal data set would cover produced output power of either a wind turbine or at least a wind farm close to Luleå. Weather data for training and forecasting should be from Luleå as well. Challenge to use training data from another wind farm and then apply the trained model to another wind farm has been researched, e.g. Hu et al. [99]. Using trained model in another location might be a larger problem.

Many different models have been developed to forecast wind power. Performance of the same model is different in different geographical locations since weather conditions vary. No model is the best one everywhere. Several models using the same local weather variables should be compared in order to find the optimal one. Support vector regression and various neural network structures would be worth studying with Luleå weather data for comparison.

Models are also developed for different time-horizons. Model, which is performing well in very short-term or short-term does not necessarily be the best choice for long-term application. Question is what kind of time-ranges do different applications of a data center need, and then select suitable models for each application.

Many AI implementations for a data center could be studied. Data center's interest is to minimize electricity cost, or to check automatically when to buy or sell extra energy from or to main grid. If sustainability and carbon neutrality are targets, then data center would like to maximize usage of renewable energies. AI could also help to reduce downtime, optimize server functionality, and monitor equipment for failures. See Section 2.2.3 for further information.

8. CONCLUSION

Sustainability has made several trends relevant in technology. The ones related to data centers are renewable energy sources, reduction of energy consumption, carbon neutrality, green microgrids, and reduction of water consumption. Data centers are increasingly using renewables, such as wind and solar energy. RISE's ICE data center has already solar panels and is now studying impact of adding wind turbine into their microgrid. Purpose of this thesis was to develop a machine learning model to forecast wind power production for the data center.

Wind power forecasting has many applications in technology. Renewable energy sources are intermittent, so accurate forecasting reduces need for additional balancing of energy and reserve power in an electricity grid. Wind turbines can be controlled, and their maintenance breaks scheduled to suitable time points. Data center has several applications. Renewable energy can be reserved from market for next hour or next day to maximize its use. Forecasting from 30 min to 6 hours ahead allows job scheduling to optimize usage of renewables and reduce power consumption. Data center may target to minimize electricity cost or maximize usage of renewables for lower greenhouse gas emissions. Smart microgrid based on artificial intelligence is the way to implement the applications.

Several different forecasting models have been developed for wind power: physical, statistical, and combined ones with both physical and statistical approaches. The data available supports choosing a statistical model. Random forest regression was used in the research. It is a widely used model for regression problems; it can handle large data sets and large amount of missing data.

Two open data sets have been used in the research. Indian data covered variables from one wind turbine, so it helped to develop forecasting model. Only one year of usable data was available. Two years would have covered a seasonal variability better, and thus, trained the model to be more accurate. Developed model forecasted output power well. Swedish data set is from a competition to estimate total wind power production in Sweden and had to be applied to approximate one wind turbine in Luleå. Output power of Luleå price region was averaged, and location for the simulation was chosen to be near Luleå. As expected, accuracy of forecasting with Swedish data was reasonable, but approximations done reduced it.

The developed model was applied for RISE's ICE data center. A validation has been done, but a final testing will take place in RISE's simulation environment. In general, data for wind power forecasting from northern Sweden is not openly available. In addition, no scientific articles from the area were noticed. Study with competition data gave understanding, which variables are significant in northern Sweden and about their relative importances. Using two data sets from different geographical locations proved that climate has a major impact on performance of the trained model. Thus, it is reasonable to use the trained model in locations with similar weather conditions only.

9. REFERENCES

[1] Geng, H. (2015) *Data center handbook*. John Wiley & Sons, Inc.

[2] Miltz, K. (2021) Global data center IP traffic 2012 – 2020 by data center type. Statista forecast.

[3] International Energy Agency, IEA (2020) *Data Centres and Data Transmission Networks*, IEA, Paris https://www.iea.org/reports/data-centres-and-data-transmission-networks

[4] Gao, J. (2014) Machine learning applications for data center optimization. Google research article on computer science.

[5] Alphabet Inc. Annual report 2020. 10-K form for the fiscal year ended December 31, 2020.

[6] Google Environmental Report 2019

[7] Apple Inc. Annual report 2020. 10-K form

[8] Facebook Inc. Annual report 2020. 10-K form

[9] All in: Staying the course on our commitment to sustainability, Amazon sustainability report, June 2020.

[10] Amazon.com Inc. Annual report 2020

[11] Tuulivoimalla katettiin noin 10% Suomen sähkönkulutuksesta vuonna 2020 (in Finnish). Suomen tuulivoimayhdistyksen tiedote. 2. 2. 2021 https://tuulivoimayhdistys.fi/ajankohtaista/tiedotteet/tuulivoimalla-katettiin-noin-10-suomen-sahkonkulutuksesta-vuonna-2020

[12] Tuulivoimaennusteita. Suomen tuulivoimayhdistys. https://tuulivoimayhdistys.fi/tietoa-tuulivoimasta-2/tietoa-tuulivoimasta/tuulivoima-suomessa-ja-maailmalla/tuulivoimaennusteita

[13] Selvitystyö Suomen tuulivoimasta – visio 2030 (in Finnish). Suomen tuulivoimayhdistys & Gasum Portfolio Services Oy, 2020, 32 p. https://tuulivoimayhdistys.fi/

[14] Christensen, J. D., Therkelsen, J., Georgiev, I., & Sand, H. (2018). *Data centre opportunities in the Nordics: An analysis of the competitive advantages*. Nordic Council of Ministers.

[15] Heinermann, J. P. (2016). *Wind power prediction with machine learning ensembles* (Doctoral dissertation, Universität Oldenburg).

[16] Holttinen, H., Miettinen, J. J., & Sillanpää, S. (2013). Wind power forecasting accuracy and uncertainty in Finland.

[17] Trading. European power exchange Nord Pool AS https://www.nordpoolgroup.com/trading/

[18] Fingrid: Balancing energy and energy prices. https://www.fingrid.fi/en/electricity-market/reserves_and_balancing/balancing-energy-and-balancing-capacity-markets/

[19] Pohjoismaiden suurin sähkövarasto Viinamäen tuulipuistoon. (in Finnish) Fingridin tiedote 30. 8. 2019 https://www.fingridlehti.fi/pohjoismaiden-suurin/

[20] Pohjoismaiden suurin akkuvarasto nousee Yllikkälään. (in Finnish) Fingridin tiedote 17. 6. 2020 https://www.fingridlehti.fi/pohjoismaiden-suurin-akkuvarasto-nousee-yllikkalaan/

[21] Virtual power plant (n.d.) Next Kraftwerke report https://www.next-kraftwerke.com/vpp/virtual-power-plant

[22] Tuulivoimaloiden hyötysuhde nousee ja tuotantokustannukset laskevat ((in Finnish). Tuulivoimayhdistyksen tiedote 12. 6. 2019. https://tuulivoimayhdistys.fi/ajankohtaista/tiedotteet/tuulivoimaloiden-hyotysuhde-nousee-ja-tuotantokustannukset-laskevat

[23] Tuulivoimatuotanto talvella (n.d.) (in Finnish). Suomen tuulivoimayhdistys.

[24] Turkia, V., Holttinen, H. (2011) Statistics of wind power production. Annual report 2011. VTT technology 74.

[25] Energiaviraston tiedote 18. 6. 2020 (in Finnish) https://energiavirasto.fi/-/aurinkosahkon-tuotantokapasiteetti-jatkoi-kasvuaan-vuonna-2019-vuosikasvua-64-prosenttia

[26] Rong, H., Zhang, H., Xiao, S., Li, C., & Hu, C. (2016). Optimizing energy consumption for data centers. *Renewable and Sustainable Energy Reviews*, *58*, 674-691.

[27] Li, Y., Orgric, A. Menaud, J. (2017) Balancing the use of batteries opportunistic scheduling policies for maximizing renewable energy consumption in a cloud data center. In *2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)* (pp. 408-415). IEEE.

[28] Asmus, P. (2017) Data centers and advance microgrids. Meeting resiliency, efficiency, and sustainability goals through smart and cleaner power infrastructure. White paper. Schneider Electric.

[29} Joshi, N. (2020) Exploring the impact of AI in the data center. Forbes May 31st 2019. https://www.forbes.com/sites/cognitiveworld/2019/05/31/exploring-the-impact-of-ai-in-the-data-center/

[30] Auringonsäteilyn määrä Suomessa. (n.d.) (in Finnish) Motivan artikkeli. https://www.motiva.fi/ratkaisut/uusiutuva_energia/aurinkosahko/aurinkosahkon_perusteet/auringonsateilyn_maara_suomessa

[31] Foley, A. M., Leahy, P. G., Marvuglia, A., & McKeogh, E. J. (2012). Current methods and advances in forecasting of wind power generation. *Renewable Energy*, *37*(1), 1-8.

[32] Jung, J., & Broadwater, R. P. (2014). Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews*, *31*, 762-777.

[33] Siebert, N. (2008). *Development of methods for regional wind power forecasting* (Doctoral dissertation, École Nationale Supérieure des Mines de Paris).

[34] Nielsen, T., Madsen, H. & Nielsen, H. (2001) Zephyr – the prediction models.

[35] Lange, M., Focken, U., Heinermann D. (2002) Regional wind power prediction with risk control. World wind energy conference, Berlin

[36] Python windpowerlib library https://pypi.org/project/windpowerlib/

[37] Milligan, M., Schwartz, M. N., & Wan, Y. H. (2003). *Statistical wind power forecasting for US wind farms* (No. NREL/CP-500-35087). National Renewable Energy Lab., Golden, CO (US).

[38] Liu, H., Tian, H. Q., Chen, C., & Li, Y. F. (2010). A hybrid statistical method to predict wind speed and wind power. *Renewable energy*, *35*(8), 1857-1861.

[39] Wang, M. D., Qiu, Q. R., & Cui, B. W. (2012, July). Short-term wind speed forecasting combined time series method and arch model. In *2012 International Conference on Machine Learning and Cybernetics* (Vol. 3, pp. 924-927). IEEE.

[40] Miranda, M. S., & Dunn, R. W. (2006, June). One-hour-ahead wind speed prediction using a Bayesian methodology. In *2006 IEEE Power Engineering Society General Meeting* (pp. 6-pp). IEEE.

[41] Fan, S., Liao, J. R., Yokoyama, R., Chen, L., & Lee, W. J. (2009). Forecasting the wind generation using a two-stage network based on meteorological information. *IEEE Transactions on Energy Conversion*, *24*(2), 474-482.

[42] Wang, J., Hu, J., Ma, K., & Zhang, Y. (2015). A self-adaptive hybrid approach for wind speed forecasting. *Renewable Energy*, *78*, 374-385.

[43] Scikit: Feature importances with forests of trees. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

[44] Breiman, L. (2001) Random forests. *Machine learning*, *45*(1), 5-32.

[45] Lahouar, A., & Slama, J. B. H. (2017). Hour-ahead wind power forecast based on random forests. *Renewable energy*, *109*, 529-541.

[46] Zhou, Z., Li, X., & Wu, H. (2016, December). Wind Power Prediction based on Random Forests. In *2016 4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016)* (pp. 352-356). Atlantis Press.

[47] Fischer, A., Montuelle, L., Mougeot, M., & Picard, D. (2017). Statistical learning for wind power: A modeling and stability study towards forecasting. *Wind Energy*, *20*(12), 2037-2047.

[48] Kramer, O., Gieseke, F., & Satzger, B. (2013). Wind energy prediction and monitoring with neural computation. *Neurocomputing*, *109*, 84-93.

[49] Kramer, O., & Gieseke, F. (2011). Short-term wind energy forecasting using support vector regression. In *Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011* (pp. 271-280). Springer, Berlin, Heidelberg.

[50] Kramer, O., & Gieseke, F. (2011, July). Analysis of wind energy time series with kernel methods and neural networks. In *2011 Seventh International Conference on Natural Computation* (Vol. 4, pp. 2381-2385). IEEE.

[51] Zeng, J., & Qiao, W. (2011, March). Support vector machine-based short-term wind power forecasting. In *2011 IEEE/PES Power Systems Conference and Exposition* (pp. 1-8). IEEE.

[52] Botha, N., & van der Walt, C. M. (2017, November). Forecasting wind speed using support vector regression and feature selection. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)* (pp. 181-186). IEEE.

[53] Li, L. L., Zhao, X., Tseng, M. L., & Tan, R. R. (2020). Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm. *Journal of Cleaner Production*, *242*, 118447.

[54] Meraihi, Y., Ramdane-Cherif, A., Acheli, D., & Mahseur, M. (2020). Dragonfly algorithm: a comprehensive review and applications. *Neural Computing and Applications*, 1-22.

[55] A. Zameer, A. Khan, S. G. Javed, et al. (2015) Machine learning based short term wind power prediction using a hybrid learning model. *Computers & Electrical Engineering*, 45:122{133, 2015}

[56] Manero Font, J. (2020). *Deep learning architectures applied to wind time series multi-step forecasting* (Doctoral dissertation, Universitat Politècnica de Catalunya).

[57] Khan, M., Liu, T., & Ullah, F. (2019). A new hybrid approach to forecast wind power for large scale wind turbine data using deep learning with TensorFlow framework and principal component analysis. *Energies*, *12*(12), 2229.

[58] Ghadi, M. J., Gilani, S. H., Afrakhte, H., & Baghramian, A. (2014). A novel heuristic method for wind farm power prediction: A case study. *International Journal of Electrical Power & Energy Systems*, *63*, 962-970.

[59] Grassi, G., & Vecchio, P. (2010). Wind energy prediction using a two-hidden layer neural network. *Communications in Nonlinear Science and Numerical Simulation*, *15*(9), 2262-2266.

[60] Díaz–Vico, D., Torres–Barrán, A., Omari, A., & Dorronsoro, J. R. (2017). Deep neural networks for wind and solar energy prediction. *Neural Processing Letters*, *46*(3), 829-844.

[61] Wang, H. Z., Li, G. Q., Wang, G. B., Peng, J. C., Jiang, H., & Liu, Y. T. (2017). Deep learning based ensemble approach for probabilistic wind power forecasting. *Applied energy*, *188*, 56-70.

[62] Kitajima, T., Yasuno, T., & Sori, H. (2013). Study on output prediction system of wind power generation using complex-valued neural network with multipoint GPV data. *IEEJ transactions on electrical and electronic engineering*, *8*(1), 33-39.

[63] Liu, Z., Gao, W., Wan, Y. H., & Muljadi, E. (2012, September). Wind power plant prediction by using neural networks. In *2012 IEEE energy conversion congress and exposition (ECCE)* (pp. 3154-3160). IEEE.

[64] Bilal, B., Ndongo, M., Adjallah, K. H., Sava, A., Kébé, C. M., Ndiaye, P. A., & Sambou, V. (2018, February). Wind turbine power output prediction model design based on artificial neural networks and climatic spatiotemporal data. In *2018 IEEE International Conference on Industrial Technology (ICIT)* (pp. 1085-1092). IEEE.

[65] Negnevitsky, M., Johnson, P. L., & Santoso, S. (2007). Short term wind power forecasting using hybrid intelligent systems.

[66] Sideratos, G., & Hatziargyriou, N. D. (2007). An advanced statistical method for wind power forecasting. *IEEE Transactions on power systems*, *22*(1), 258-265.

[67] Damousis, I. G., Alexiadis, M. C., Theocharis, J. B., & Dokopoulos, P. S. (2004). A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation. *IEEE Transactions on Energy Conversion*, *19*(2), 352-361.

[68] Jursa, R., & Rohrig, K. (2008). Short-term wind power forecasting using evolutionary algorithms for the automated specification of artificial intelligence models. *International Journal of Forecasting*, *24*(4), 694-709.

[69] Azeem, A., Fatema, N., & Malik, H. (2018). k-NN and ANN based deterministic and probabilistic wind speed forecasting intelligent approach. *Journal of Intelligent & Fuzzy Systems*, *35*(5), 5021-5031.

[70] Zhang, Y., & Wang, J. (2016). K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting. *International Journal of forecasting*, *32*(3), 1074-1080.

[71] Mori, H., & Kurata, E. (2008, November). Application of gaussian process to wind speed forecasting for wind power generation. In *2008 IEEE International Conference on Sustainable Energy Technologies* (pp. 956-959). IEEE.

[72] Kou, P., & Gao, F. (2012, June). Sparse heteroscedastic Gaussian process for shortterm wind speed forecasting. In *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[73] Yan, J., Li, K., Bai, E. W., Deng, J., & Foley, A. M. (2015). Hybrid probabilistic wind power forecasting using temporally local Gaussian process. *IEEE Transactions on sustainable energy*, *7*(1), 87-95.

[74] Pinson, P., Ranchin, T., & Kariniotakis, G. (2004, March). Short-term wind power prediction for offshore wind farms Evaluation of Fuzzy-Neural network based models. In *Global Windpower Conference* (pp. CD-ROM).

[75] Kariniotakis, G., Moussafir, J., Buty, D., Usaola, J., et al. (2004) Towards next generation short-term forecasting of wind power – The ANEMOS project. Global WindPower 2004 Conference, March 2004, Chigaco, USA

[76] Wang, X., Guo, P., & Huang, X. (2011). A review of wind power forecasting models. *Energy procedia*, *12*, 770-778.

[77] Monteiro, C. Bessa, R., Miranda, V., Botterud, A., Wang, J. & Gonzelmann, G. (2009) Wind Power Forecasting: State-of-the-Art 2009. Argonne national laboratory report ANL/DIS-10-1.

[78] Alea Soft web page
https://aleasoft.com/products-and-services/wind-energy-forecasting/

[79] DNV LG forecasting services
https://www.dnv.com/services/forecaster-introduction-3848

[80] Gonzalez, G., Diaz-Guerra, B., Soto, F., Lopez, S., Sanchez, I., Usaola, J., ... & Lobo, M. G. (2004). SIPREÓLICO-Wind power prediction tool for the Spanish peninsular power system. *Proceedings of the CIGRÉ 40th general session and exhibition, París, France*.

[81] Nielsen, T. S., Madsen, H., & Tofting, J. (2000, July). WPPT a tool for on-line wind power prediction. In *Proc. Int. Energy Agency Expert Meeting Wind Forecasting Techniques*.

[82] Barbosa de Alencar, D., de Mattos Affonso, C., Limão de Oliveira, R. C., Moya Rodriguez, J. L., Leite, J. C., & Reston Filho, J. C. (2017). Different models for forecasting wind power generation: Case study. *Energies*, *10*(12), 1976.

[83] Giebel, G., Draxl, C., Brownsword, R., Kariniotakis, G., & Denhard, M. (2011). The state-of-the-art in short-term prediction of wind power. A literature overview. EU project Anemos.

[84] Rodrigues, A. () EPREV – A wind power forecasting tool for Portugal

[85] INT-WPFS web page
http://www.nj-int.com/solution/info/18.html

[86] LocalPRED: Accurate wind forecasts for wind energy (n.d.) CENER report.

[87] Möhrlen, C., & Jørgensen, J. U. (2006). Forecasting wind power in high wind penetration markets using multi-scheme ensemble prediction methods. *Proceedings of DEWEK 2006*.

[88] Cali, U., Lange, B., Dobschinski, J., Kurt, M., Moehrlen, C., & Ernst, B. (2008, May). Artificial neural network based wind power forecasting using a multi-model approach. In *Proceedings of the 7th International Workshop on Large-Scale Integration of Wind Power and on Transmission Networks for Offshore Wind Farms, Madrid (ES)*.

[89] Bhaskarpandit, S. (2017 - 2018) Wind power forecasting – two-and-half years data for a wind turbine. Kaggle online community of data scientists and machine learner practitioners.
https://www.kaggle.com/theforcecoder/wind-power-forecasting

[90] EEM20 wind power forecasting competition (2020) 17th International conference on the European energy market. 16 – 18 September, Stockholm, Sweden
https://eem20.eu/forecasting-competition/

[91] Sweden outline map illustrations & vectors. Royalty free.
https://www.dreamstime.com/illustration/sweden-outline-map.html

[92] Scikit. Random Forest Regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[93] Scikit: Partial dependence plots https://scikit-learn.org/stable/modules/partial_dependence.html#partial-dependence-plots

[94] Molnar, C. (2020) Interpretable machine learning – a guide for making black box models explainable. Lulu. 318 p. Also found in https://christophm.github.io/interpretable-ml-book/intro.html

[95] RISE Infrastructure and Cloud research & test Environment (ICE) data center. https://www.ri.se/en/ice-datacenter

[96] Siltala, M. (2020) Simulating data center cooling systems: data-driven and physical modeling methods. Master's theses. Aalto University.

[97] Brännvall, R., Siltala, M., Gustafsson, J., Sarkinen, J., Vesterlund, M., & Summers, J. (2020, June). Edge: Microgrid data center with mixed energy storage. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems* (pp. 466-473).

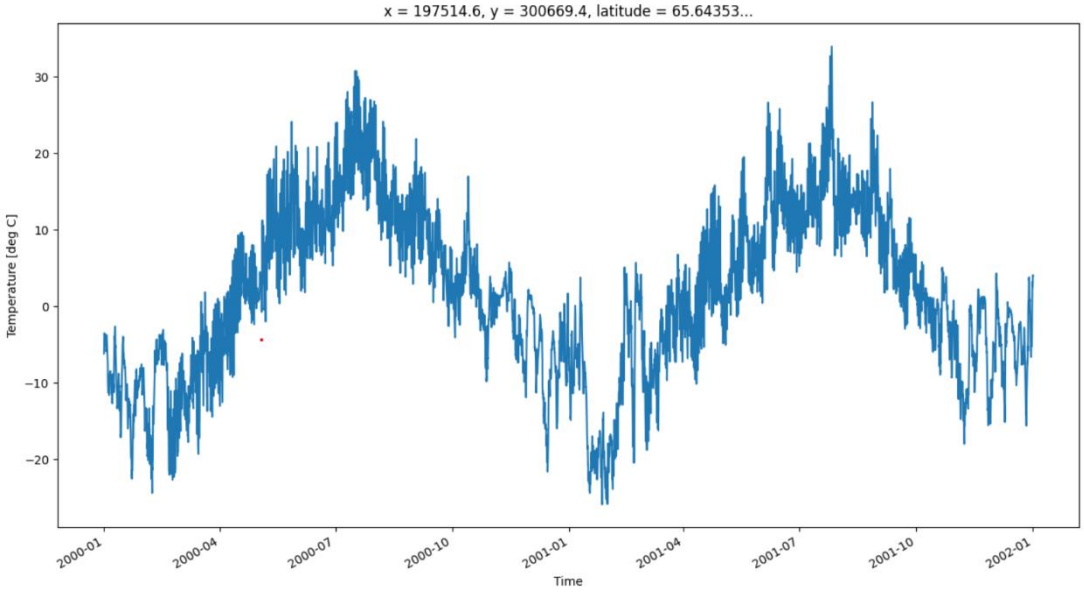
[98] Swedish meteorological and hydrological institute. 10 day weather forecasting data for Luleå. http://www.smhi.se/en/weather/sweden-weather/ten-day-forecast/q/Lule%C3%A5/604490

[99] Hu, Q., Zhang, R., & Zhou, Y. (2016). Transfer learning for short-term wind speed prediction with deep neural networks. *Renewable Energy*, *85*, 83-95.

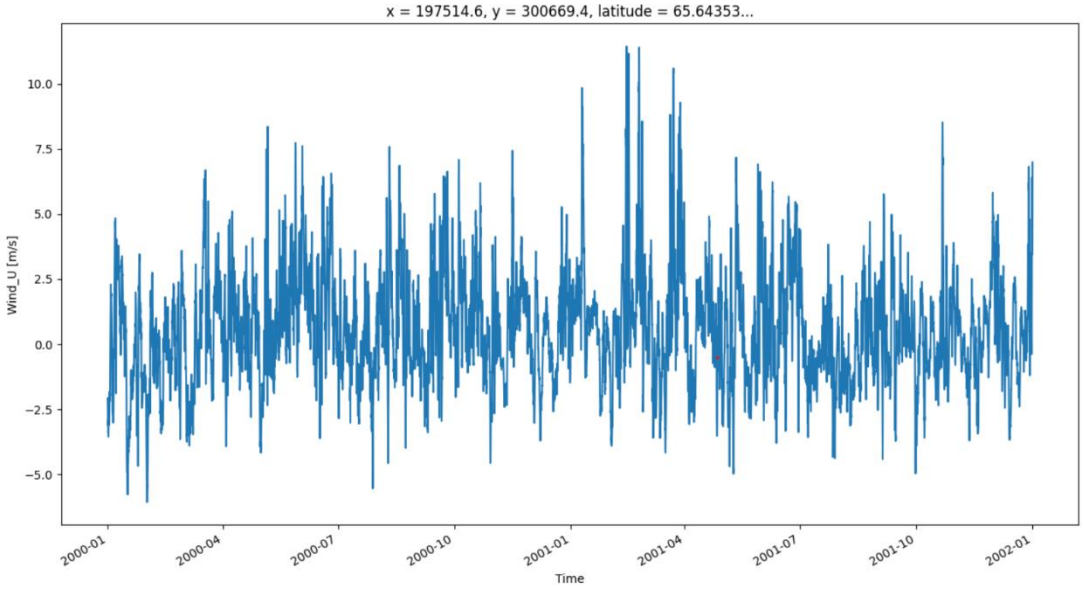
10. APPENDICES

Appendix 1. Weather data, years 2000 and 2001. Location 66.07N, 17.16E.

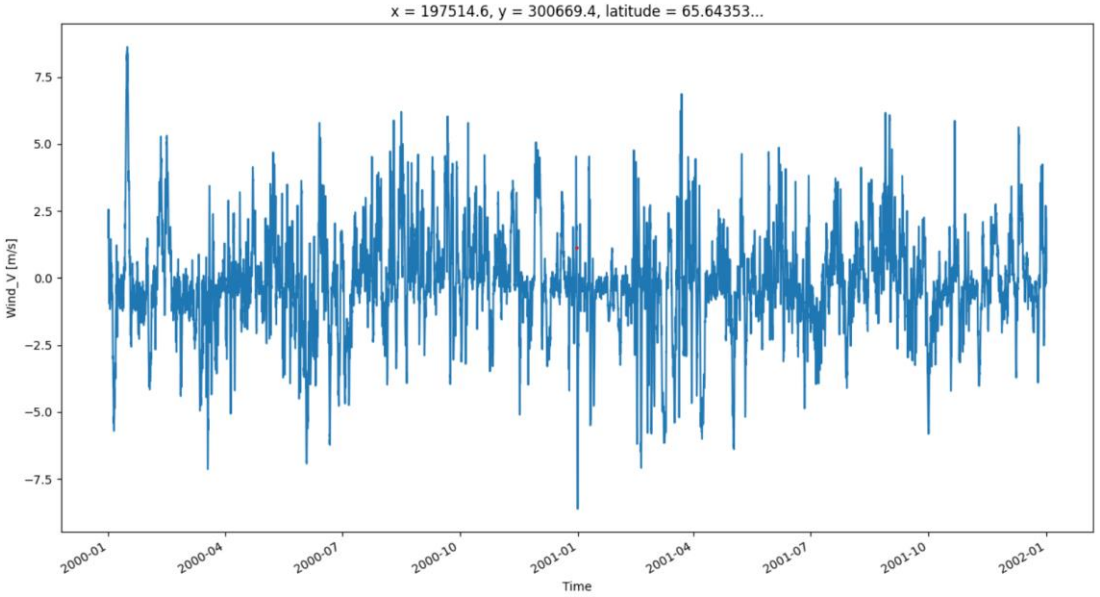
Temperature [°C]



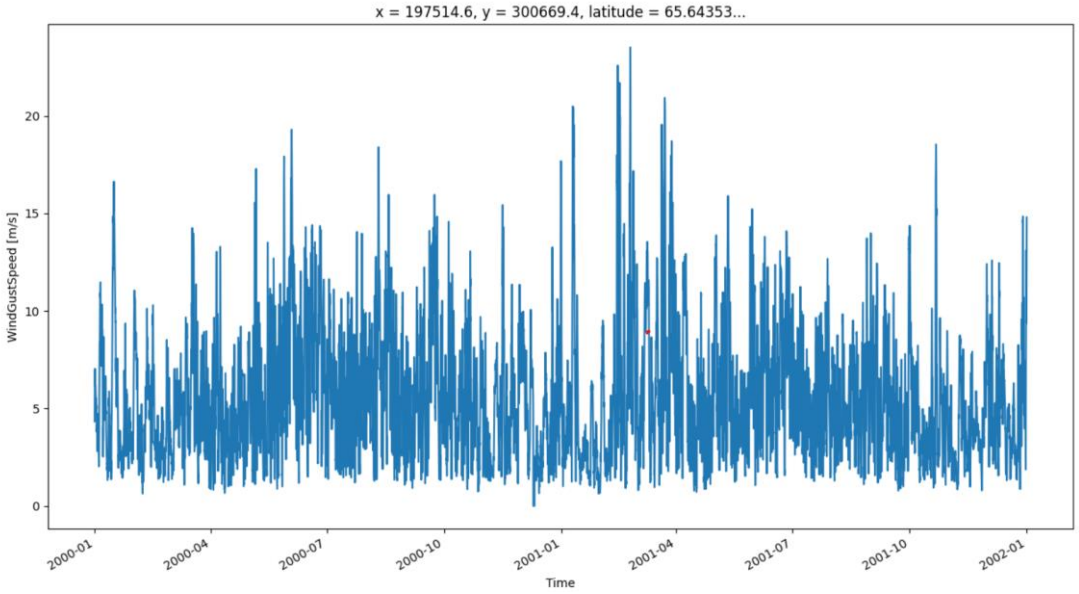
Zonal wind speed [m/s]



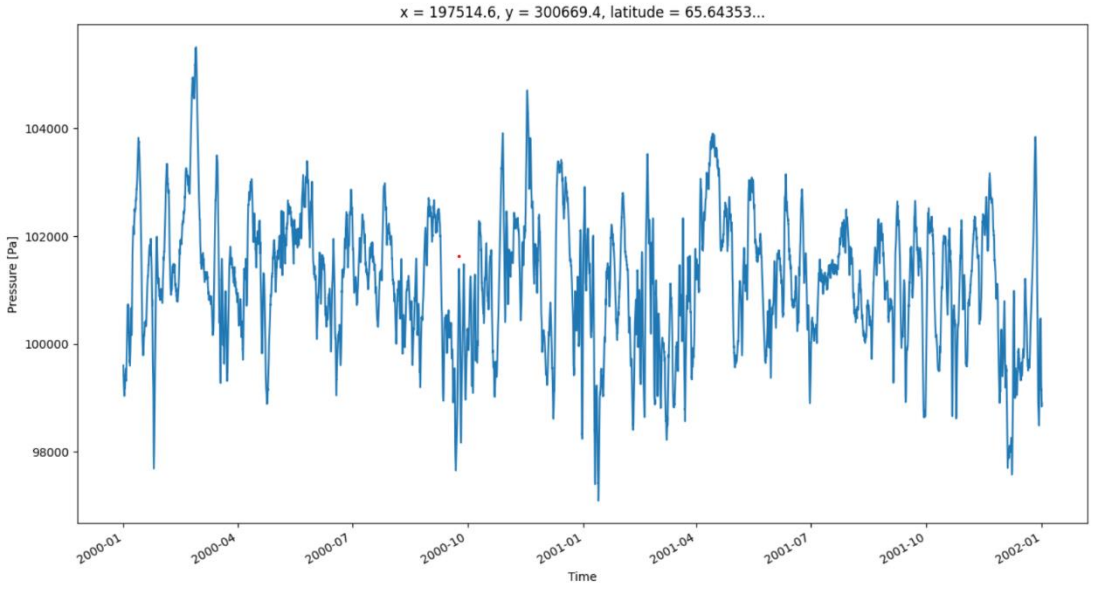
Meridional wind speed [m/s]



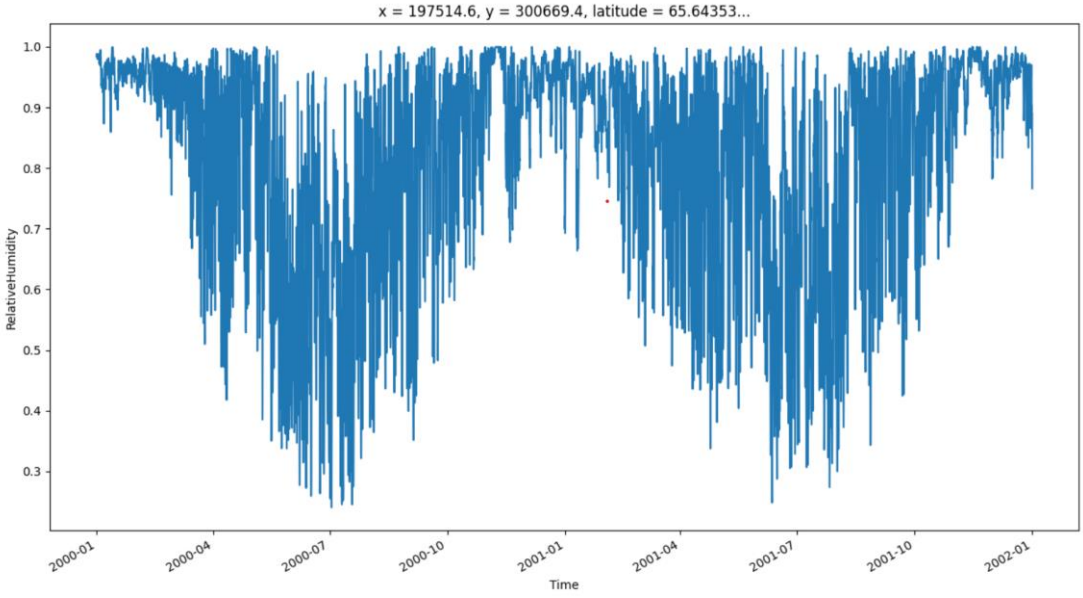
Wind gust speed [m/s]



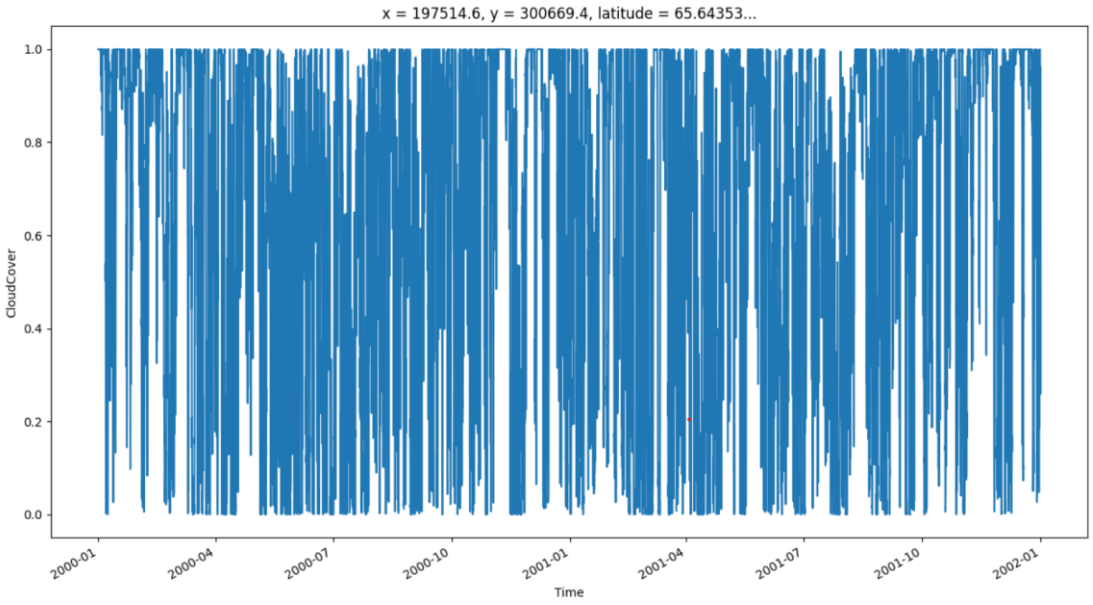
Pressure [Pa]



Relative humidity

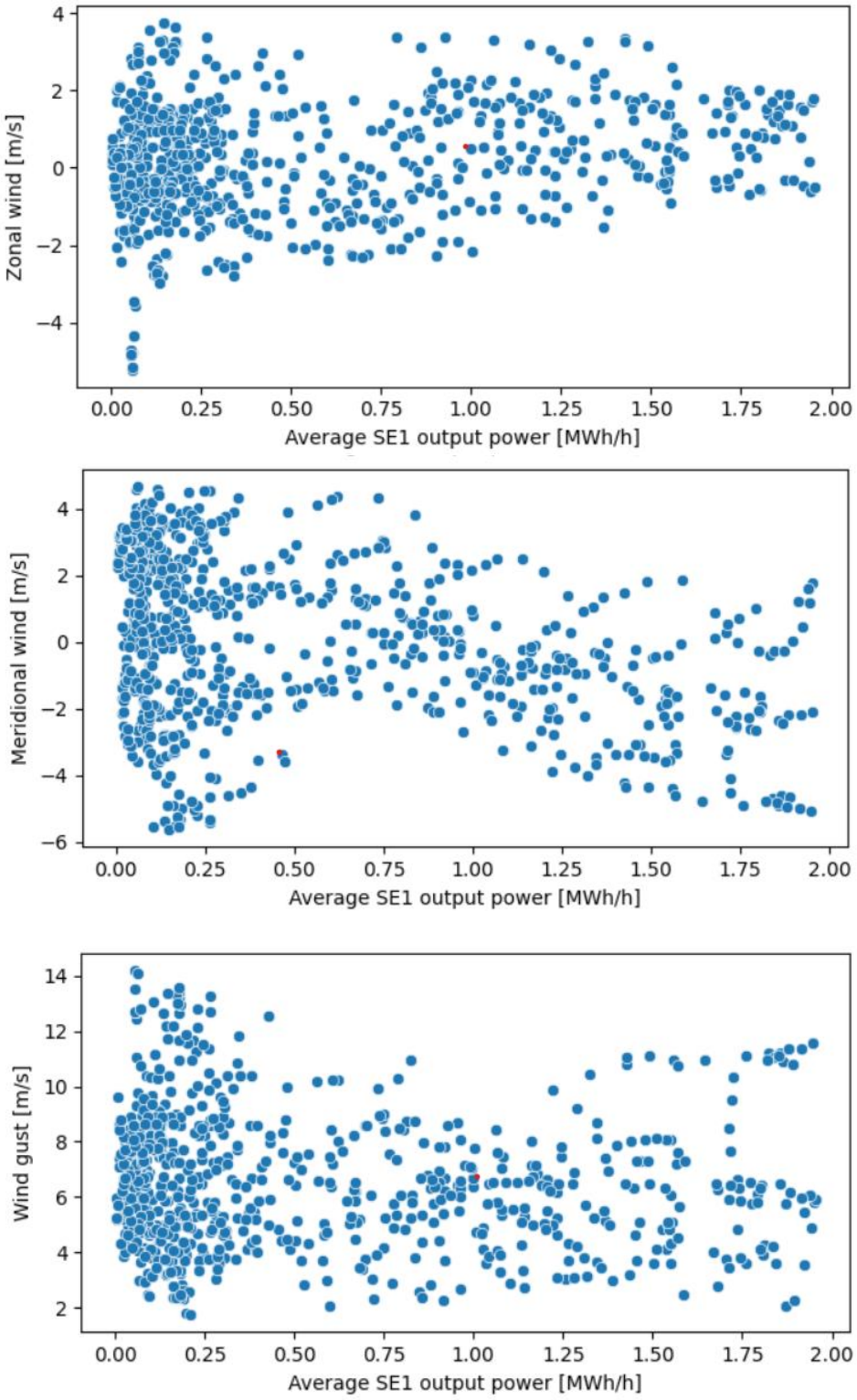


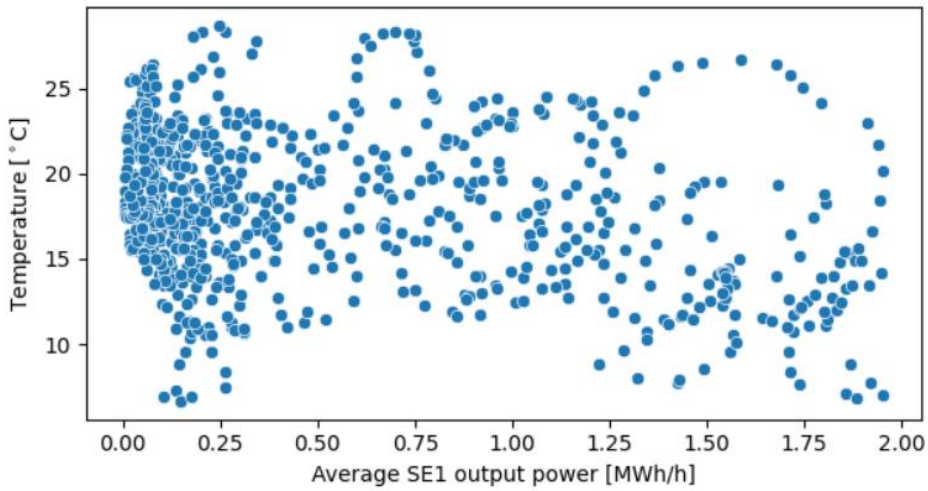
Cloud cover

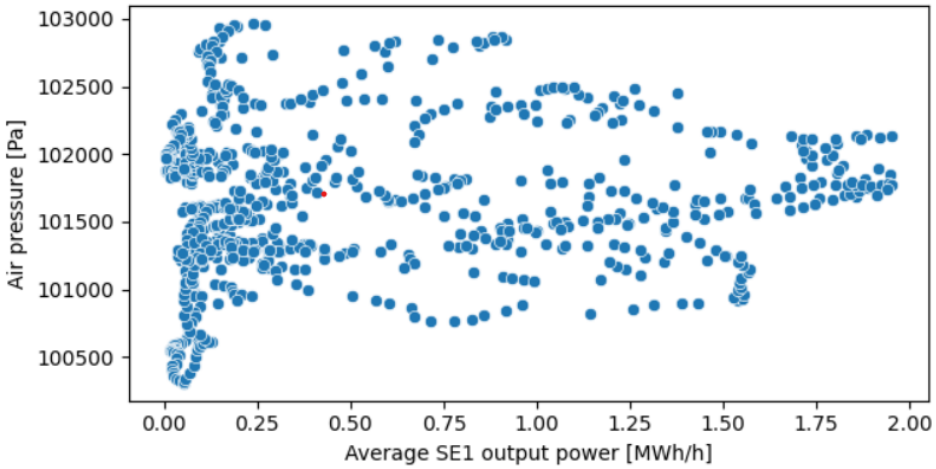

x = 197514.6, y = 300669.4, latitude = 65.64353...
CloudCover
1.0
0.8
0.6
0.4
0.2
0.0
2000-01
2000-04
2000-07
2000-10
2001-01
2001-04
2001-07
2001-10
2002-01
Time

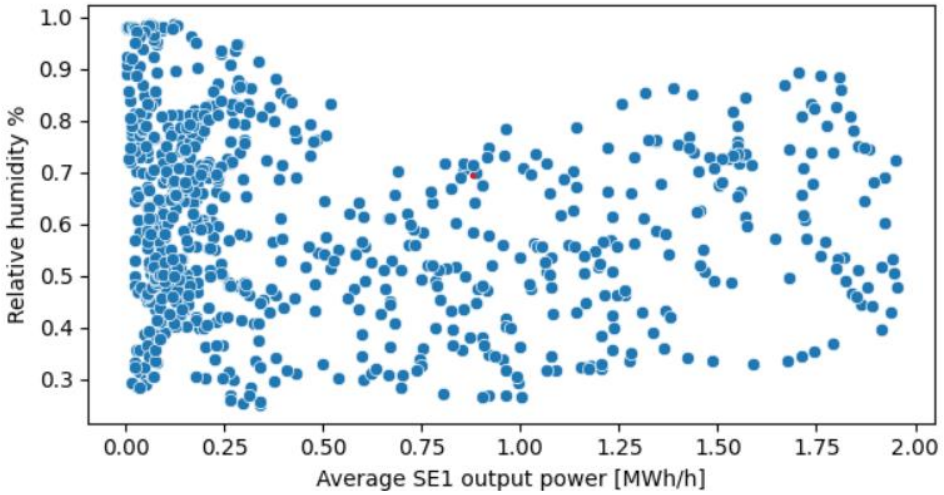
Appendix 2. Relationships between averaged production power and input variables, January 2000, and July 2000. Location 66.07N, 17.16E.

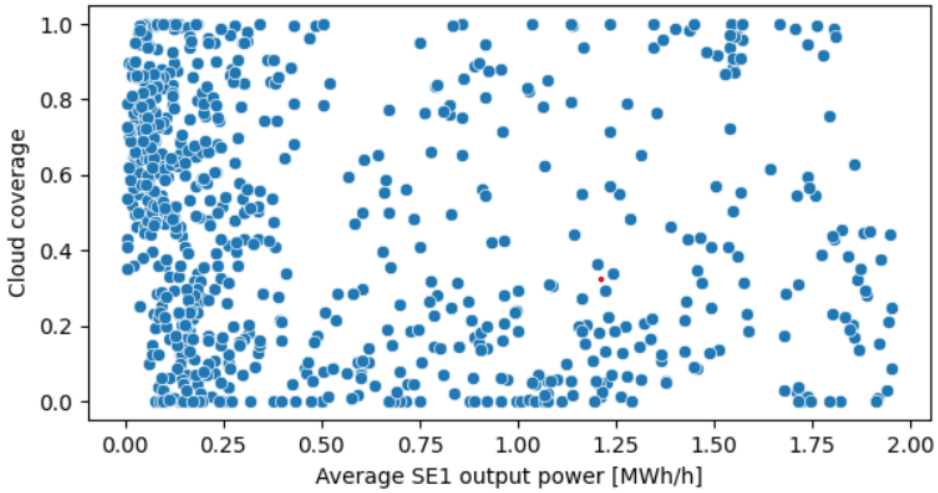
January 2000



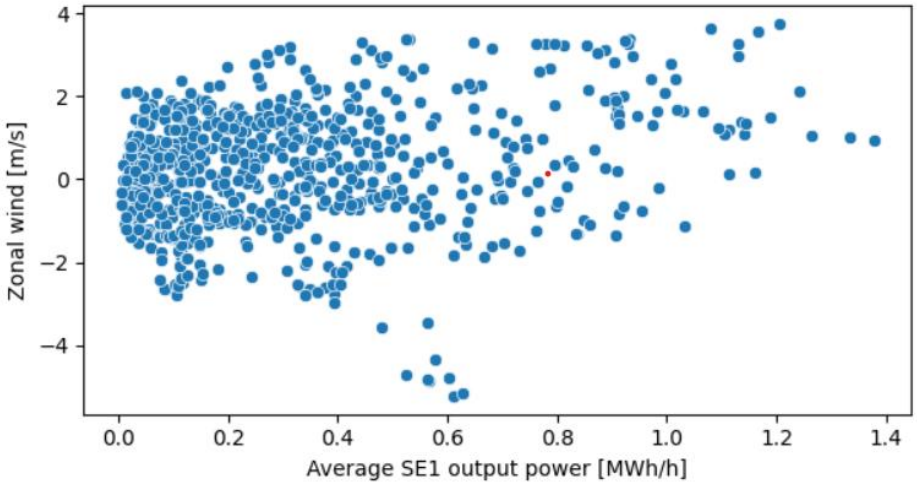

Temperature [°C]
25
20
15
10
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Average SE1 output power [MWh/h]

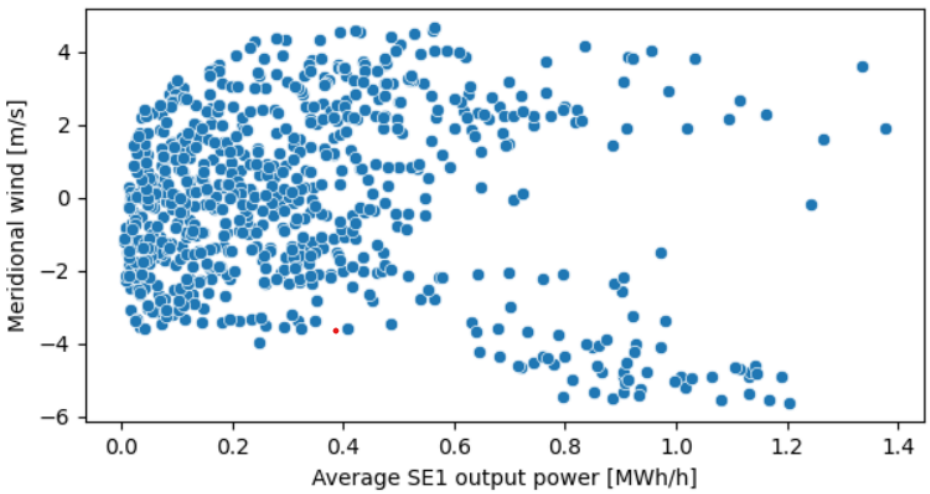

Air pressure [Pa]
103000
102500
102000
101500
101000
100500
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Average SE1 output power [MWh/h]

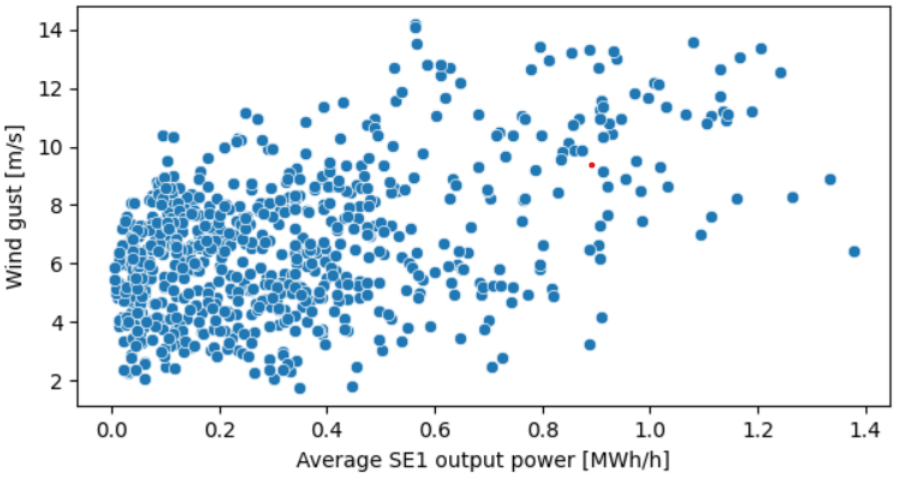

Relative humidity %
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Average SE1 output power [MWh/h]

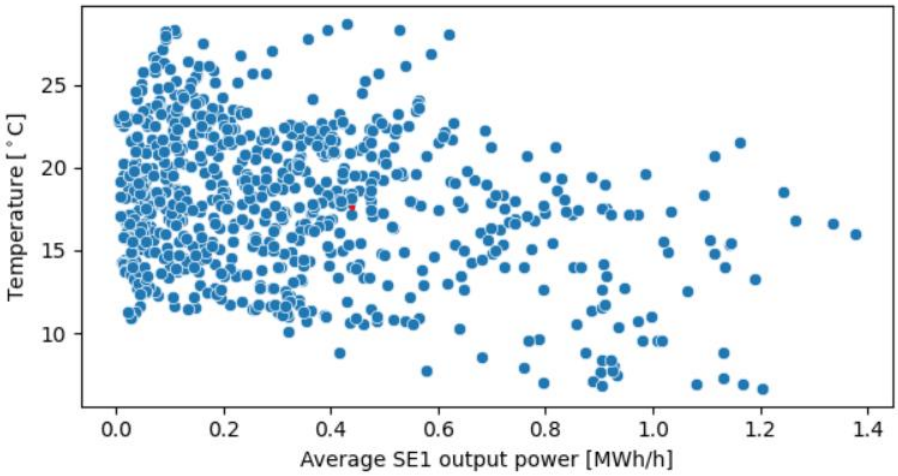

Cloud coverage
Average SE1 output power [MWh/h]

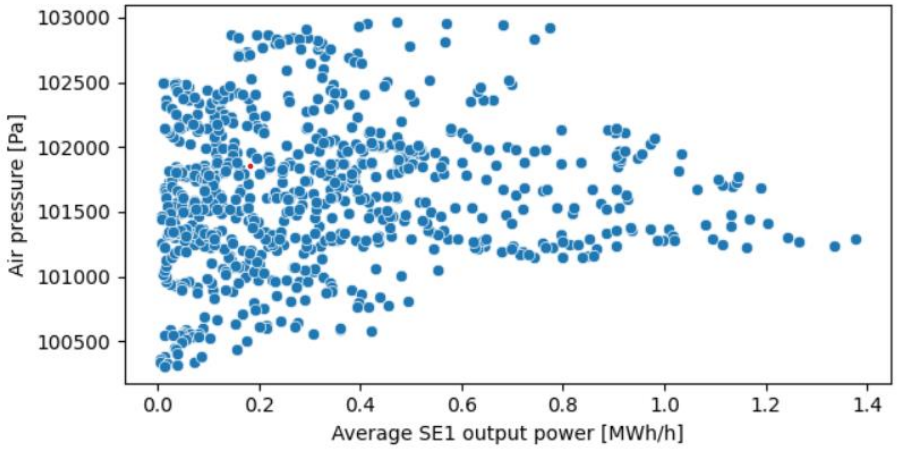
July 2000

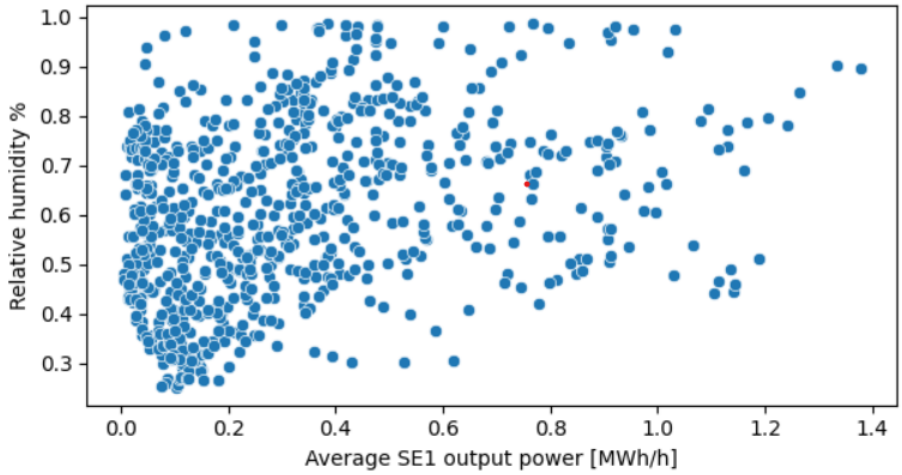

Zonal wind [m/s]
Average SE1 output power [MWh/h]

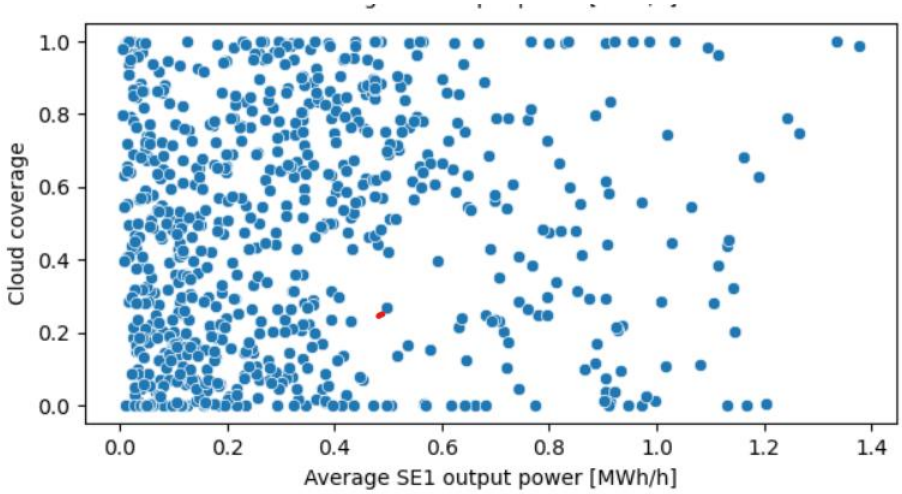

Meridional wind [m/s]
Average SE1 output power [MWh/h]


Wind gust [m/s]
14
12
10
8
6
4
2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Average SE1 output power [MWh/h]


Temperature [° C]
25
20
15
10
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Average SE1 output power [MWh/h]

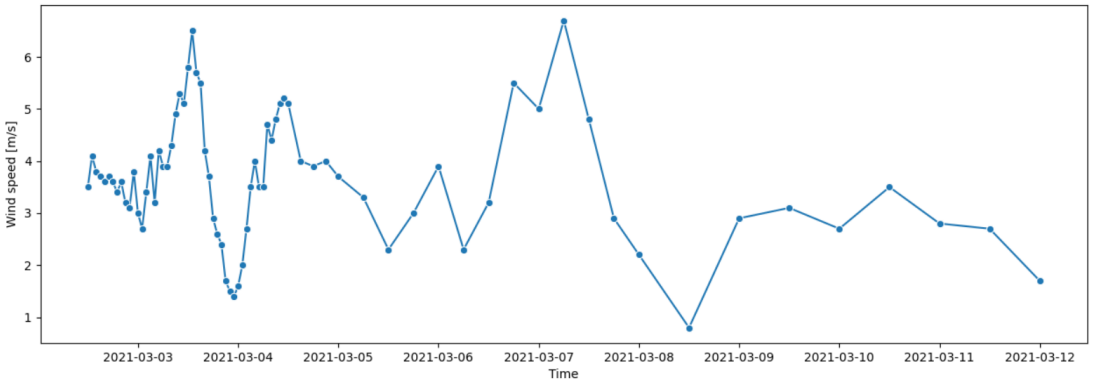

Air pressure [Pa]
103000
102500
102000
101500
101000
100500
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Average SE1 output power [MWh/h]


Relative humidity %
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Average SE1 output power [MWh/h]

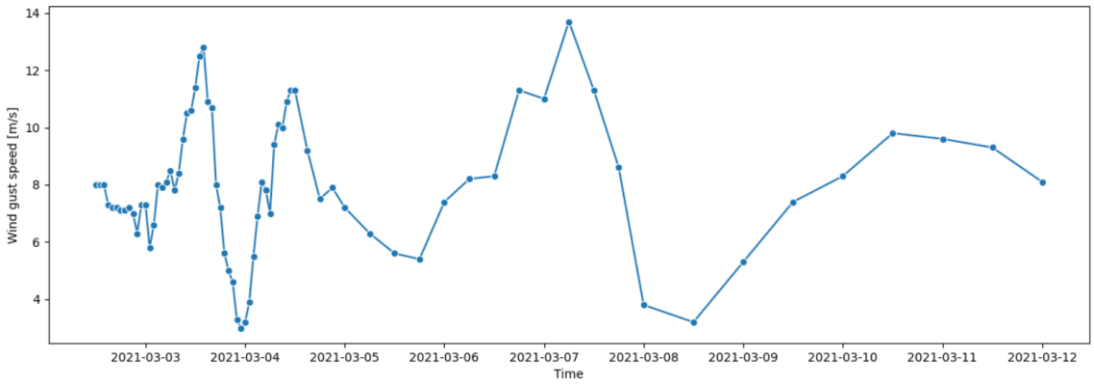

Cloud coverage
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Average SE1 output power [MWh/h]

Appendix 3. Input variables for RISE test code. Additionally, calculated zonal wind speed and meridional wind speed. Location RISE, Luleå, Sweden

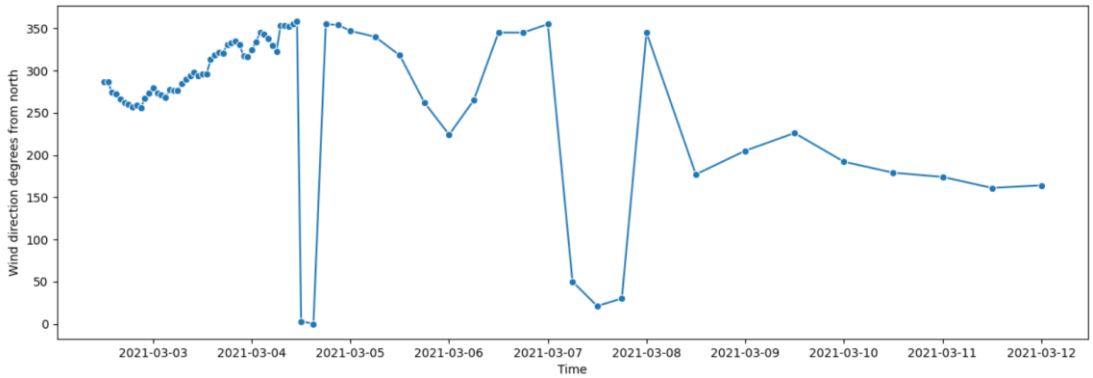
Wind speed [m/s]



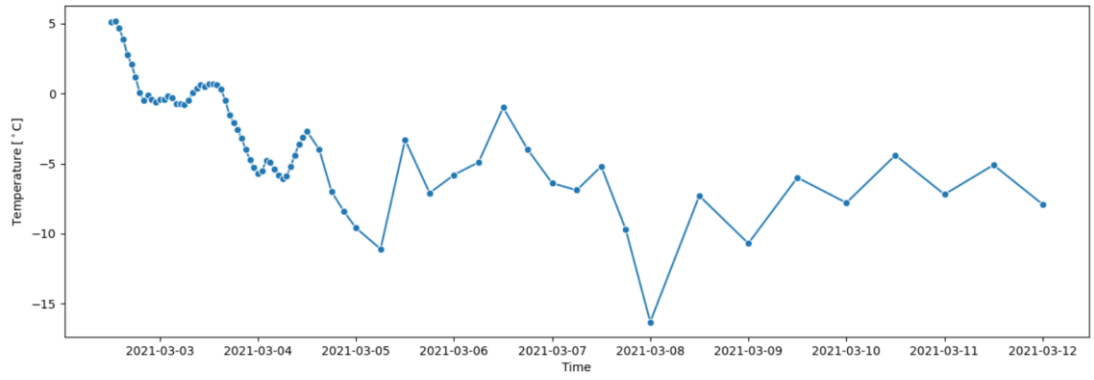
Wind gust speed [m/s]



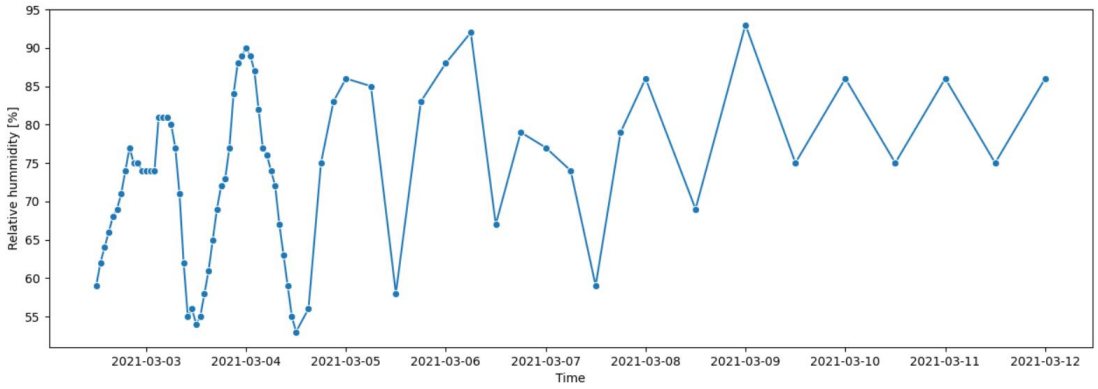
Wind direction [degrees from north]



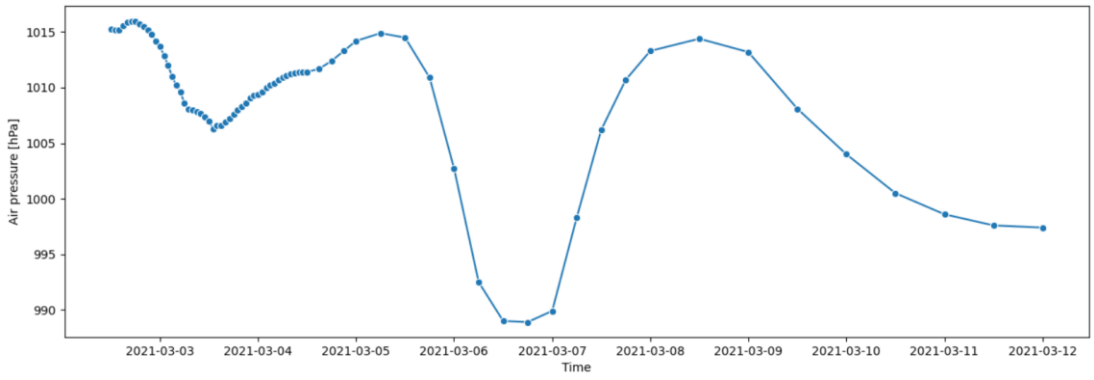
Temperature [°C]



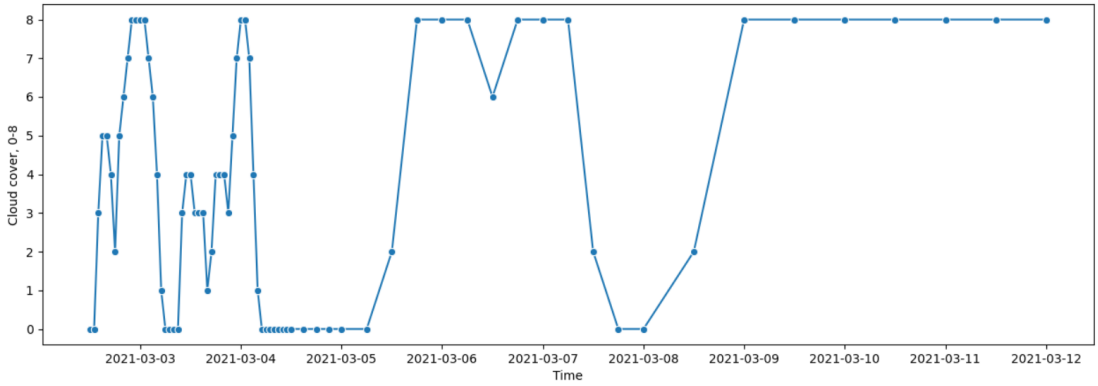
Relative humidity



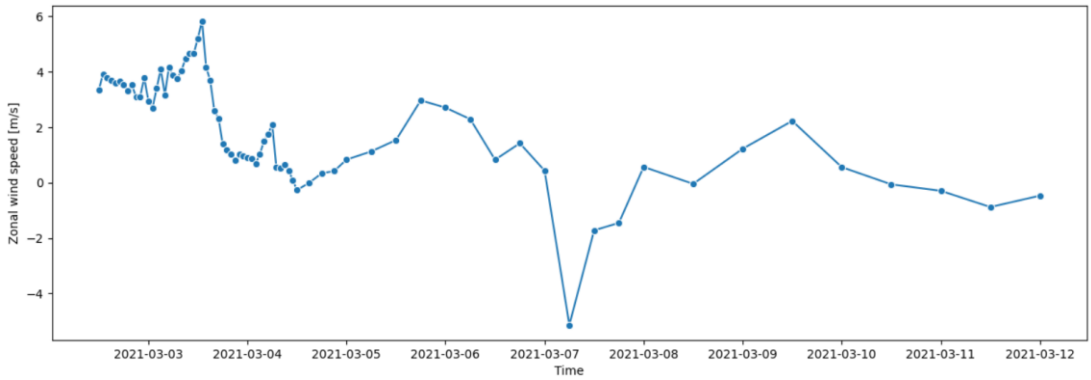
Air pressure [hPa]



Cloud cover



Calculated zonal wind speed



Calculated meridional wind speed